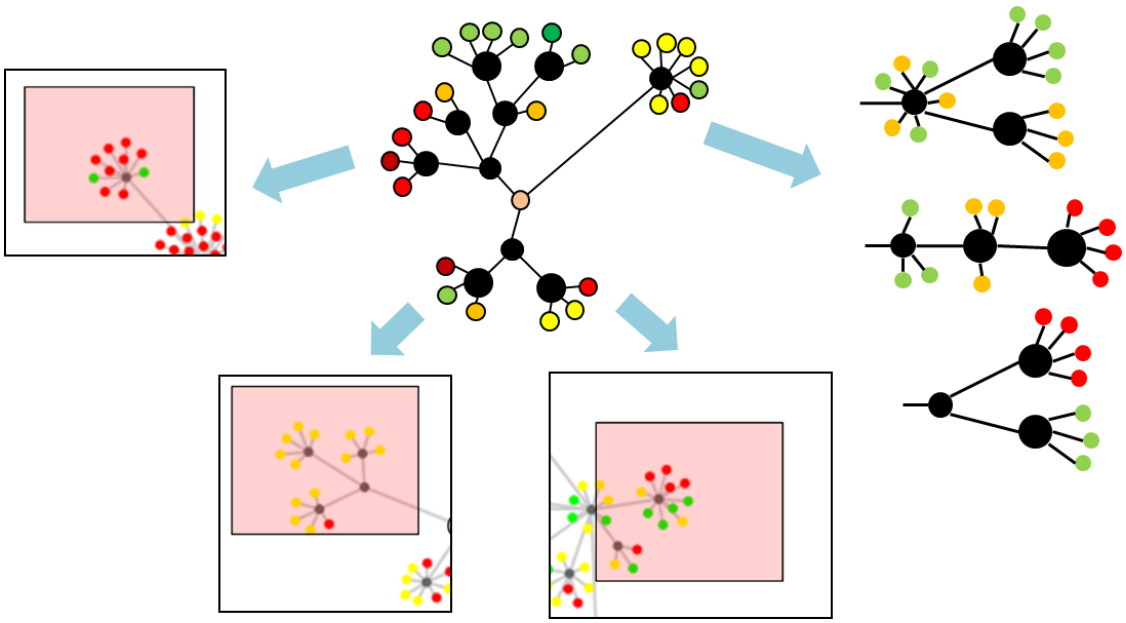
Technische Universität Braunschweig

inSARa:
Hierarchische Netzwerke zur Analyse, Visualisierung und Vorhersage von Struktur-Aktivitäts-Beziehungen



Von der Fakultät für Lebenswissenschaften

der Technischen Universität Carolo-Wilhelmina zu Braunschweig

zur Erlangung des Grades einer

Doktorin der Naturwissenschaften

(Dr. rer. nat.)

genehmigte

D i s s e r t a t i o n

von Elgin Sabrina Wollenhaupt
aus Goslar

1. Referent: Prof. Dr. Knut Baumann
2. Referent: Prof. Dr. Conrad Kunick
3. Referent: Prof. Dr. Gisbert Schneider
eingereicht am: 03.02.2014
mündliche Prüfung (Disputation) am: 20.06.2014

Druckjahr 2014

Vorveröffentlichungen der Dissertation

Teilergebnisse aus dieser Arbeit wurden mit Genehmigung der Fakultät für Lebenswissenschaften, vertreten durch den Mentor der Arbeit, in folgenden Beiträgen vorab veröffentlicht:

Publikationen

(1) Wollenhaupt, S.; Baumann, K. inSARa: Intuitive and Interactive SAR Interpretation by Reduced Graphs and Hierarchical MCS-based Network Navigation. *J. Chem. Inf. Model.* (Manuskript in Revision)

Tagungsbeiträge

Vorträge

(1) Wollenhaupt, S., Baumann, K.: inSARa: Intuitive Single-Target (Large-Scale) SAR Interpretation and Multi-Target Cross-Reactivity Analysis. 9th German Conference on Chemoinformatics, Fulda, Germany (2013).

(2) Wollenhaupt, S., Baumann, K.: inSARa: Analysis of Structure-Activity Relationships and Target Prediction. 1. MINAS Symposium, Warberg, Germany (2013).

(3) Wollenhaupt, S., Baumann, K.: inSARa: Intuitive automatisierte Analyse von Struktur-Aktivitäts-Beziehungen in großen Datensätzen. Seminar der Pharmazeutischen Institute SS 2013, Braunschweig, Germany (2013).

(4) Wollenhaupt, S., Baumann, K.: inSARa: Intuitive and Interactive SAR Interpretation by Hierarchical MCS-based Network Navigation. European CCG UGM and Conference 2013, Amsterdam, the Netherlands (2013).

Posterbeiträge

(1) Wollenhaupt, S., Baumann, K.: inSARa: Enabling Intuitive and Interactive (Large-Scale) SAR Analysis by Reduced Graphs and Hierarchical MCS-based Network Navigation. 6th Joint Sheffield Conference on Cheminformatics, Sheffield, UK (2013).

(2) Wollenhaupt, S., Baumann, K.: inSARa Networks for Ligand-based Analysis of Target Similarities. EuroCUP VI, Santpoort, the Netherlands (2013).

(3) Wollenhaupt, S., Baumann, K.: Combination of Fingerprints and MCS-based (inSARa) Networks for Structure-Activity-Relationship Analysis. 8th German Conference on Chemoinformatics, Goslar, Germany (2012).

(4) Wollenhaupt, S., Baumann, K.: inSARa: SAR Interpretation by Network Navigation. DPHG-Jahrestagung, Greifswald, Germany (2012).

(5) Wollenhaupt, S., Baumann, K.: Comparison of Fingerprints and MCS-based (INSARA) Structure-Activity-Relationship Networks. 3rd Strasbourg Summer School on Chemoinformatics, Strasbourg, France (2012).

(6) Wollenhaupt, S., Baumann, K.: INSARA: Tackling Large-Scale SAR Analysis with Reduced Graphs and MCS-based Networks. OpenEye CUP XIII, Santa Fe, New Mexico, USA (2012).

(7) Wollenhaupt, S., Baumann, K.: INSARA: A New Method for the Analysis and Visualization of Structure-Activity-Relationships. 7th German Conference on Chemoinformatics, Goslar, Germany (2011).

Folgende Publikation wurde während des Promotionsverfahrens zum Druck angenommen:

(1) Wollenhaupt, S.; Baumann, K. inSARa: Intuitive and Interactive SAR Interpretation by Reduced Graphs and Hierarchical MCS-based Network Navigation. *J. Chem. Inf. Model.* **2014**, *54* (6), 1578–1595.

Allen, die in meinem Herzen sind.

Meinen Eltern, Großeltern und Esthi.

“To be or not to be - that is the question…”

W. Shakespeare

Danksagung

Die vorliegende Arbeit entstand in der Zeit von Januar 2011 bis Dezember 2013 am Institut für Medizinische und Pharmazeutische Chemie der Technischen Universität Carolo-Wilhelmina zu Braunschweig unter der Betreuung von

Herrn Professor Dr. Knut Baumann,

dem ich an dieser Stelle für das Überlassen dieses unglaublich vielseitigen und faszinierenden Themas, für seine gute Betreuung, als auch für die vielen konstruktiven und inspirierende Gespräche ganz herzlich danken möchte. Lieber Knut, ich bin dir sehr dankbar für deine große Unterstützung (v.a. auch in den finalen Wochen) und all die Möglichkeiten, die du mir während meiner Promotionszeit gegeben hast, sowie die Freiheit, die du mir bei der Bearbeitung meines Themas und der Gestaltung meiner Arbeit gelassen hast! Ich danke dir für drei sehr lehrreiche Jahre, die mich persönlich sehr geprägt haben und eine wichtige Lebenserfahrung waren!

Mein besonderer Dank gilt ebenso

Herrn Professor Dr. Conrad Kunick

für die Übernahme der Tätigkeit des Zweitgutachter.

Ebenso danke ich herzlich

Frau Professor Dr. Ute Wittstock

für die Übernahme der Tätigkeit des Drittprüfers und die Leitung der Prüfungskommission.

Zudem danke ich allen gegenwärtigen und ehemaligen Mitgliedern des Arbeitskreises Baumann für die angenehme Zusammenarbeit und Unterstützung in den vergangenen drei Jahren, hier insbesondere *Dr. Sebastian G. Rohrer, Dr. Markus Kossner, Dr. Jan Dreher, Dr. Stephanie Ludewig, Dr. Magnus Matz, Désirée Baumann, Shantheya Balasupramaniam, Waldemar Klingspohn, Miriam Mathea, Dr. Juan Manuel Pallicer Santana* und *Dr. Florian Kölling*. Ich wünsche Euch allen weiterhin viel Erfolg für den weiteren beruflichen, als auch den privaten Lebensweg!

Für die konstruktive Zusammenarbeit im Rahmen der Betreuung des 5. Semesters (Praktikum „Arzneistoffanalytik“) möchte ich mich ebenfalls bei allen beteiligten Kollegen, sowie Chefs und natürlich allen Studenten, die ich betreuen durfte, bedanken. Danke, dass ich zur Expertin für „Verschmutzungen“ werden durfte! Das Arzneibuch ist mir sehr ans Herz gewachsen! Danke v.a. für die wichtige Unterstützung in der finalen „Endspurt“-Phase!

Dr. Hans-Otto Burmeister und Dr. Lutz Preu möchte ich für all ihr Engagement im Rahmen der Weiterbildung zum Fachapotheker für pharmazeutische Analytik danken und dass sie bei Fragen jeder Art immer ein verlässlicher Ansprechpartner waren!

Neben den bereits genannten Personen möchte ich mich natürlich auch *bei allen weiteren Mitarbeitern des Instituts, insbesondere bei meiner „guten Fee von nebenan“ Britta Thomas,* für die gute Zusammenarbeit bedanken.

Ein spezieller Dank geht an *Miriam Mathea*! Miri, ich danke dir von Herzen für deine Unterstützung und dass man sich immer auf dich verlassen kann! Danken möchte an dieser Stelle auch *Volker*, *Shanthy* und *Waldemar* sowie an *Juanma und Therese*, die ebenfalls das letzte Jahr meiner Promotionszeit mitgeprägt haben! Schade, dass die gemeinsame Zeit nur so kurz war! Ich wünsche euch nur das Beste für die Zukunft!

Ein besonderes Dankeschön geht auch an *Dr. Florian Kölling*!
Flo, danke für die vielen tausend (Dienst-)Reise-Kilometer, die wir zusammen zurückgelegt haben – einmal quer durch Europa und bis ans Ende der Welt! Diese und die unzähligen Stunden, die ich mich dabei vor Lachen nicht halten konnte, werde ich immer in besonderer Erinnerung behalten! Vielen lieben Dank für all deine Unterstützung und Motivation! Ich wünsche dir alles erdenklich Gute für die Zukunft und den weiteren Lebensweg!

Danken möchte ich auch allen *„Glücksbärchis“ aus der Pharmazeutischen Biologie,* die mit ihrer immerwährenden guten Laune mir immer ein Lächeln auf die Lippen zaubern konnten! Ein ganz besonderes Dankeschön geht an *Anja Böhme* und *Marion Wiggermann*! Vielen Dank, dass ihr immer da wart und dass wir nach dem Studium auch diese Lebensphase zusammen erfolgreich „durchleiden“ und meistern durften! Ich wünsche euch von ganzem Herzen alles, alles Gute und ganz viel Erfolg für die Zukunft!

Auch bei *Tobias Fiesel* möchte ich mich an dieser Stelle ganz herzlichst bedanken!
Tobi, danke für jegliche Unterstützung, LOLs und aufbauende Worte! Auch dir wünsche ich alles erdenklich Liebe und ganz viel Erfolg für die Zukunft!

Es gibt viele Momente und schöne, unvergessene Erlebnisse, die mich immer an euch erinnern werden. Ich werde euch alle sehr, sehr vermissen! Ich hoffe aber und wünsche mir von ganzem Herzen, dass es kein Abschied für immer ist und dass sich unsere (Lebens-) Wege noch lange nicht trennen werden!

So, und nun am Ende noch die *wichtigsten Menschen*, die diese Arbeit maßgeblich beeinflusst haben – wenn auch nicht unbedingt inhaltlich – und ohne die es diese Arbeit in dieser Form so nicht gegeben hätte:

Mein tiefster Dank gilt *meiner Schwester* für all ihre fortwährende Unterstützung, ihre ununterbrochene Motivation und dafür, dass sie immer für mich da war! Ich danke dir von allertiefstem Herzen für Alles! Danke, das werde ich dir niemals vergessen!

Meinen lieben Großeltern gilt ebenfalls ein ganz besonderer Dank! Danke, dass ihr mich immer gefördert und unterstützt habt!

Von allertiefstem Herzen danke ich auch *meinen geliebten Eltern*, die mich immer unterstützt haben, immer für mich da waren und immer an mich geglaubt haben! Ohne euch wäre ich heute nicht da, wo ich jetzt bin, und ohne euch würde es diese Arbeit ganz sicher nicht geben! DANKE, Mama, DANKE, Papa!

Inhaltsverzeichnis

Abkürzungsverzeichnis

Abkürzung	deutsch	englisch
AC	sprunghafte SARs	**A**ctivity **c**liff
ADME	Aufnahme, Verteilung, Metabolismus und Ausscheidung (=Pharmakokinetik)	**a**bsorption, **d**istribution, **m**etabolism and **e**xcretion
ANN	künstliche neuronale Netze	**A**rtificial **N**euronal **N**etwork
BMMSG		**B**ipartite **M**atching **M**olecular **S**eries **G**raph
BMS		**B**emis **M**urcko **S**caffold
CADD	computergestützte Arzneistoffentwicklung	**C**omputer-**a**ided **d**rug **d**esign
CAG		**C**ombinatorial **A**nalog **G**raph
CNG		**C**hemical **N**eighborhood **G**raph
CATS		**C**hemical **A**dvanced **T**emplate **S**earch
CB1	Cannabinoid Rezeptor 1	**C**anna**b**inoid receptor **1**
CDK2	Cyclin-abhängige Kinase 2	**C**yclin-**d**ependent **k**inase **2**
COX2	Cyclooxygenase 2	**C**yclo**ox**ygenase **2**
CS	gemeinsame Substruktur	**C**ommon **s**ubstructure
CSD		**C**hambridge **S**tructural **D**atabase
CSK	zyklische Gerüste	**C**yclic **sk**eletons
EC_{50}	halbmaximale effektive Konzentration	
ECFP		**E**xtended-**C**onnectivity **F**inger**p**rint
ErG	erweiterten reduzierten Graphen	**E**xtended **r**educed **G**raph
FCFP		**F**unctional-**C**lass **F**inger**p**rint
FDA		**F**ood and **D**rug **A**dministration
FEPOPS		**FE**ature **PO**int **P**harmacophor**S**
FXa	Koagulations-**F**aktor **Xa**	Coagulation **F**actor **Xa**
FP	molekularer Fingerabdruck/ Fingerprint	**F**inger**p**rint
FT	Eigenschaftsbaum	**F**eature **T**ree
GPCR	G-Protein gekoppelter Rezeptor	**G**-**p**rotein **c**oupled **r**eceptor
HBA	H-Brücken-Akzeptor	**H**ydrogen **b**ond **a**cceptor
HBAD	gemeinsame **H**-**B**rücken-**A**kzeptor und -**D**onor-Eigenschaft	
HBD	**H**-**B**rücken-**D**onor	**H**ydrogen **b**ond **d**onor

hERG		**h**uman **E**ther-à-go-go-**R**elated **G**ene
HTS	Hochdurchsatz-Screening	**H**igh-**t**hroughput **s**creening
IC_{50}	halbmaximale inhibitorische Konzentration	
InCHI		IUPAC **In**ternational **C**hemical **I**dentifier
inSARa		**i**ntuitive **n**etworks for **S**tructure-**A**ctivity **R**elationship **A**nalysis
ISAC	Identifizierung von strukturbasierten sprunghaften SARs	**I**dentification of **S**tructure-based **A**ctivity **C**liffs
IT-Assay	Isolierter Target Assay	
IUBMB		**I**nternational **U**nion of **B**iochemistry and **M**olecular **B**iology
IUPAC		**I**nternational **U**nion of **P**ure and **A**pplied **C**hemistry
IUPHAR		**I**nternational **U**nion of Basic and Clinical **Phar**macology
K_d	Dissoziationskonstante	
K_i	Gleichgewichtsdissoziationskonstante des Inhibitors	
***k*NN**	k-nächste-Nachbarn	**k**-**N**earest-**N**eighbors
LASSO Graph		**LA**yered **S**keleton−**S**caffold **O**rganization **Graph**
MCS	maximal gemeinsame Substruktur / maximal gemeinsamer Subgraph	**M**aximum **c**ommon **s**ubstructure / **m**aximum **c**ommon **s**ubgraph
MCIS	maximal gemeinsamer induzierter Subgraph	**M**aximum **c**ommon **i**nduced **s**ubgraph
MCES	maximal gemeinsamer Kanten-Subgraph	**M**aximum **C**ommon **E**dge **S**ubgraph
MLR		**M**ultiple **L**inear **R**egression
MMP	zusammenpassendes Molekülpaar	**M**atched **M**olecular **P**air
MMPA	MMP-Analyse	**M**atched **M**olecular **P**air **A**nalysis
MMS	zusammenpassende Molekülserie	**M**atching **M**olecular **S**eries
MOE		**M**olecular **O**perating **E**nvironment
MST	Minimaler Spannbaum	**M**inimum **S**panning **T**ree
mtCAG		**m**ulti-**t**arget **CAG**
mtSAR		**m**ulti-**t**arget **SAR**
NI	**n**egativ **i**onisierbar	**n**egative **i**onizable
NSG		**N**etwork-like **S**imilarity **G**raph
P38		MAP Kinase **p38** alpha
PDB		**P**rotein **d**ata **b**ank

PI	**p**ositiv **i**onisierbar	**p**ositive **i**onizable
pIC_{50}	negativ dekadischer Logarithmus des IC_{50}-Wertes	
pK_i	negativ dekadischer Logarithmus des K_i-Wertes	
PLS		**P**artial **L**east **S**quares Regression
PRESS	Quadratwurzel aus dem vorhergesagten Fehler der Summe der Abweichungsquadrate	**PR**edicted **E**rror **S**um of **S**quares
QSAR	quantitative Struktur-Aktivitäts-Beziehung	**Q**uantitative **S**tructure-**A**ctivity **R**elationship
QSPR	quantitative Struktur-Eigenschafts-Beziehung	**Q**uantitative **S**tructure-**P**roperty **R**elationship
RASCAL		**RA**pid **S**imilarity **CAL**culation
RG	reduzierter Graph	**R**educed **G**raph
RCS		**R**educed **c**yclic **s**keletons
ROCS		**R**apid **O**verlay of **C**hemical **S**tructures
RMSEP	Quadratwurzel des mittleren quadrierten Fehlers der Datenvorhersage	**R**oot **M**ean **S**quared **E**rror of **P**rediction
R^2	quadrierter, multipler Korrelationskoeffizient der Datenvorhersage (Bestimmtheitsmaß)	
SALI		**S**tructure-**A**ctivity **L**andscape **I**ndex
SAR	Struktur-Aktivitäts-Beziehung	**S**tructure-**A**ctivity **R**elationship
SARI		**SAR I**ndex
SAS Map		**S**tructure-**A**ctivity **S**imilarity **Map**
SEA		**S**imilarity **E**nsemble **A**pproach
SHED		**SH**annon **E**ntropy **D**escriptor
SIFt		**S**tructural **I**nteraction **F**ingerprin**t**
SMARTS		**Sm**iles **Ar**bitrary **T**arget **S**pecification
SMILES		**S**implified **M**olecular **I**nput **L**ine **E**ntry **S**ystem
SOM	selbstorganisierende Karte	**S**elf-**O**rganizing **M**ap
SOSA	selektive Optimierung von UAWs	**S**elective **O**ptimization of **S**ide **A**ctivities
SPP	Ähnlichkeitsprinzip	**S**imilarity **P**roperty **P**rinciple
SPT		**S**imilarity-**P**otency **T**ree
SSSR	kleinster Satz an kleinsten Ringen	**S**mallest **S**et of **S**mallest **R**ings
SVM		**S**upport **V**ector **M**achine

Tc	Tanimoto-Koeffizient	**T**animoto **c**oefficient
THR		**Thr**ombin
UAW	unerwünschte Arzneimittel-Wirkung	
UniProt		**Uni**versal **Prot**ein Database
VS	Virtuelles Screening	**V**irtual **S**creening
WOMBAT		**WO**rld of **M**olecular **B**ioAc**T**ivity
ZB-Assay	**Z**ell-**b**asierter **Assay**	
ZBG	**Z**ink-**b**indende **G**ruppe	

I. Einleitung und Grundlagen

1. SAR-Information: Ein wichtiger Baustein auf dem Weg zum rationalen Wirkstoffdesign

Seit 1950 wurden mehr als 1200 Arzneistoffe von der FDA zugelassen.[1] Trotz dieser großen Zahl an bereits verfügbaren Wirkstoffen, ist das Auffinden neuer Wirkstoffmoleküle aus den nachfolgenden Gründen immer noch von zentraler Bedeutung:

(1) Es gibt eine Vielzahl von Erkrankungen, die aus verschiedenen Gründen bisher nicht (ausreichend) therapiert werden können:

a) Bei den sehr seltenen Erkrankungen („orphan diseases“)[2] wie auch bei einer Reihe von Infektionskrankheiten mit global hoher Prävalenz[3], von denen v.a. die ärmeren Entwicklungsländern betroffen sind („neglected diseases“)[4], ist die geringe wirtschaftliche Rentabilität im Vergleich zur kostenintensiven Entwicklung neuer Arzneistoffe ein wichtiger Grund für fehlende Therapie-Möglichkeiten. Trotz entsprechender Förderprogramme in den vergangenen Jahren[5–6] ist großer Entwicklungsbedarf vorhanden.

b) Bei Erkrankungen mit hoher bzw. zunehmend steigender Prävalenz in den immer älter werdenden westlichen Industrienationen (z.B. Morbus Alzheimer[7] oder verschiedene Krebserkrankungen[8]) spielen andere Faktoren eine Rolle. So beruhen Krankheiten in der Regel auf einem komplexen Zusammenspiel verschiedener pathophysiologischer Prozesse[9]. Die Identifizierung eines geeigneten Targets (d.h. einer chemisch definierbaren, makromolekularen Zielstruktur, die eine spezifische Interaktion mit einem chemischen Molekül (Arzneistoff) eingeht, die mit einem entsprechenden klinischen Effekt in Bezug auf die zu behandelnde Erkrankung verknüpft ist[10]) gestaltet sich bei einigen Erkrankungen (u.a. psychische Erkrankungen[11]) extrem schwierig. Verschiedene Analysen schätzen die Zahl der Targets von verfügbaren Arzneistoffen auf mehrere Hundert.[12] Die Zahl der potentiellen Targets wird auf etwa 3000 bis 8000 geschätzt.[10] Dies zeigt, dass trotz der Einschränkung, dass nicht jede mit einer Krankheit assoziierte Zielstruktur mit kleinen Molekülen modulierbar ist[13], deutliches Potential für die Entwicklung neuer Arzneistoffe vorhanden ist.

(2) Geringe Wirkstärke, schlechte Verträglichkeit oder andere Probleme (z.B. schlechte pharmakokinetische Eigenschaften, geringe chemische Stabilität) sind weitere Gründe, die die Suche nach neuen Arzneistoffkandidaten notwendig machen.

(3) Zudem führen Resistenzen immer wieder dazu, dass vorhandene Arzneistoffe wirkungslos werden. Dies ist v.a. im Bereich der Antiinfektiva[14–16] (z.B. bei Erkrankungen wie Malaria, Tuberkulose, nosokomiale Infektionen mit multiresistenten Bakterien oder im Bereich der HIV-Therapie) und Zytostatika[17] von großer Relevanz.

Die Geschichte vieler bekannter Arzneistoffe zeigt, dass Zufall ein wichtiger Faktor im Bereich der Arzneistofffindung und -entwicklung ist.[18–20] Die 1945 mit dem Medizinnobelpreis ausgezeichnete Zufallsentdeckung der Penicilline durch Alexander Fleming 1928 ist dabei eine der bekanntesten. Ausgehend von einer zufällig mit dem Schimmelpilz Penicillum notatum kontaminierten Staphylokokken-Kultur wurde eine der wichtigsten Antibiotikaklassen gefunden, die v.a. in den Kriegszeiten vielen Menschen das Leben rettete, die ansonsten z.B. an einfachsten Wundinfektionen verstorben wären.[19, 21] Ein weiteres Beispiel für einen glücklichen Zufall stellt auch die Entdeckung des Zytostatikums Cisplatin bei der Untersuchung des Einflusses von Strom auf das Wachstumsverhalten von E. coli-Bakterien durch Barnett Rosenberg 1965 dar.[22] Platinhaltige Metallverbindungen haben die Therapie verschiedenster vorher unheilbarer maligner Tumore ermöglicht. Mittlerweile sind sie Bestandteil etwa jedes zweiten antineoplastischen Therapieschemas.[23] Aber auch der Blockbuster Viagra® wurde von der Firma Pfizer nur auf Umwegen gefunden. Der Phosphodiesterase-V-Inhibitor Sildenafil, eigentlich als Arzneistoff zur Therapie der Angina pectoris entwickelt, zeigte in der wenig aussichtsreichen klinischen Prüfung eine unerwartete UAW auf die glatte Muskulatur des Corpus cavernosum, die heute als Hauptwirkung in der Therapie der erektilen Dysfunktion gewinnbringend genutzt wird.[24] Seit der Markteinführung 1998 bis zum Ablauf des deutschen Patentes Mitte 2013 hat Pfizer mit diesem „Lifestyle“-Präparat weltweit rund 25 Milliarden US-Dollar Umsatz erzielt.[25]

Durch die schnelle Entwicklung in den letzten Jahrzehnten im Bereich der Molekularbiologie, der Bioanalytik, der Robotik und der chemischen Synthese haben sich viele neue Perspektiven für die Arzneistoffentwicklung ergeben.[26] Die Suche nach neuen Arzneistoffen kann deutlich systematischer gestaltet werden. Durch Fortschritte in der kombinatorischen Chemie kann eine große Zahl von Molekülen in kürzester Zeit automatisiert synthetisiert werden. Zudem ermöglichen modernste technologische Verfahren, täglich Tausende von diesen potentiellen Wirkstoffkandidaten in experimentellen Hochdurchsatz-Screenings (Abk. HTS, engl. high-throughput screening) auf biologische Aktivität zu testen. Die damit verbundene Hoffnung deutlich mehr neue Arzneistoffe zu finden, hat sich bisher jedoch nicht erfüllt. In Anbetracht der theoretischen Maßstäbe - die unvorstellbare Größe des chemischen Raumes für potentielle Arzneistoffkandidaten (Schätzungen zufolge zwischen 10^{60} bis 10^{100} Molekülen[27–29]) im Vergleich zu den wenigen Millionen Molekülen in den firmeninternen Substanzbibliotheken – wird schnell klar, dass auch hier Zufall eine wichtige Erfolgs-Komponente ist[30]. Die Praxiserfahrung bestätigt ebenfalls geringe Trefferraten[29]. Arzneistoffentwicklung ist noch immer vergleichbar mit „der Suche nach der Nadel im Heuhaufen“[31].

Mit Hilfe verschiedenster computergestützter Techniken („Virtuelles Screening“[32–33] und „computergestütztes Wirkstoffdesign“[34]) ist es jedoch möglich die Suche nach und die Entwicklung von neuen Wirkstoffen zu unterstützen bzw. deutlich rationaler zu gestalten. Der Vorteil dieser Techniken ist, dass sie im Vergleich zum experimentellen Screening deutlich kostengünstiger sind. Sie können jedoch die experimentelle Testung niemals ersetzen, sondern stellen eine ressourcenschonende Ergänzung dar. Sehr bekannte, therapeutisch wichtige Arzneistoffklassen stellen Ergebnisse von erfolgreichem rationalem, computergestütztem Wirkstoffdesign dar (z.B. der Carboanhydrase-Inhibitor Dorzolamid, der HIV-Protease-Inhibitor Ritonavir oder der Neuraminidase-Inhibitor Zanamivir zur Therapie der Influenza).[18]

Ein wichtiger Baustein auf dem Weg zu diesem rationalen Wirkstoffdesign ist die Kenntnis von Struktur-Aktivitäts-Beziehungen (Abk. SAR, engl. structure-activity relationship). Für die gezielte, wissensbasierte Entwicklung neuer Arzneistoffe sind daher Methoden wichtig, die in der Lage sind, diese SARs aus den verfügbaren Daten zu extrahieren. Im Vergleich zu früher, wo oftmals nur wenige Daten verfügbar waren, die von den medizinischen Chemikern manuell ausgewertet werden konnten, haben sich die Anforderungen, aber auch Möglichkeiten in den letzten Jahren deutlich verändert. Der oben beschriebene technologische Fortschritt hat zu riesigen, täglich weiter anwachsenden Mengen an Bioaktivitätsdaten geführt, die in verschiedenen Datenbanken gespeichert werden. Die nun verfügbaren Datenmengen enthalten potentiell eine große Menge an Information, die die Entwicklung neuer Moleküle deutlich beschleunigen kann. Eine schnellere Überführung in die klinische Prüfung ist nicht für den Patienten, sondern auch für den pharmazeutischen Unternehmer attraktiv, da so deutlich Entwicklungskosten eingespart werden können. Die Herausforderung, die jedoch gelöst werden muss, ist die Auswertung dieser Datenmengen. Für die Bewältigung dieser Dimensionen werden heutzutage computergestützte Verfahren benötigt, die automatisiert, die wichtigsten Informationen über SARs (im Folgenden der Einfachheit halber als „SAR-Informationen“ bezeichnet) extrahieren und möglichst anschaulich und intuitiv für den medizinischen Chemiker darstellen. Eine Vielzahl an Methoden, die dieses Problem adressieren, wurde bereits in den vergangenen Jahren entwickelt. Eine Zusammenfassung dieser Arbeiten folgt in Kapitel 2.6. Der Stein der Weisen wurde bisher jedoch noch nicht gefunden.

Das Ziel dieser Arbeit war daher, die Entwicklung einer Methode namens inSARa (Abkürzung für „intuitive networks for Structure-Activity Relationship analysis“) zur intuitiven Analyse von SARs (im Folgenden als „SAR-Analyse“ bezeichnet) großer Mengen an Bioaktivitätsdaten, die die die bereits verfügbaren Verfahren weiter ergänzt.

Die nachfolgende Arbeit ist wie folgt aufgebaut. In diesem ersten Teil der Arbeit werden zunächst einige wichtige Grundlagen, die im Zusammenhang mit dem Verständnis von Struktur-Aktivitäts-Beziehungen bzw. entwickelten Methoden zur SAR-Analyse wichtig sind, zusammengefasst (Kapitel 2 und 6). Des Weiteren werden in den Kapiteln 3, 4, 5, 7 und 8 weitere, wichtige Konzepte, die die Basis für die in dieser Arbeit entwickelte inSARa-Methode und ihre Anwendung darstellen, vorgestellt. In Teil II werden dann die inSARa-Methode, sowie der Versuchsaufbau und die Auswertung verschiedener Analysen und verschiedenen Möglichkeiten der Anwendung der Methode beschrieben. In Teil III werden schließlich Ergebnisse aus der Anwendung der inSARa-Methode sowie verschiedener Analysen gezeigt und diskutiert. In Kapitel 24 wird dann eine Zusammenfassung der Arbeit gegeben, bevor in Kapitel 25 noch ein Ausblick auf zukünftige weitere Arbeiten folgt.

2. Struktur-Aktivitäts-Beziehungen (SAR)

2.1. Definition und Bedeutung für Arzneistoffentwicklung

Die Aufklärung von Struktur-Aktivitäts-Beziehungen, also des funktionellen Zusammenhanges zwischen chemischer Struktur oder Eigenschaft und biologischer Aktivität an einer bestimmten Zielstruktur[35], stellt eine Schlüssel-Information für den Arzneistoffentwicklungs-Prozess dar. Ist bekannt, welche strukturellen Merkmale entscheidend für die Bioaktivität sind, können zielgerichtet bekannte biologisch aktive Moleküle optimiert bzw. nach strukturell neuartigen Molekülen gesucht werden.

In Abbildung 2.1 sind die verschieden Stufen des Entwicklungs-Prozesses schematisch dargestellt. Man erkennt, dass der medizinische Chemiker in vielen, v.a. aber den frühen Entwicklungs-Phasen, von Information über SARs profitieren kann. Für einen umfassenden Überblick sei auf DUFFY et al.[36] verwiesen.

SAR-Information ist z.B. bei der Auswahl von potentiellen Wirkstoffkandidaten aus der Menge im HTS aktiv getesteter Moleküle („Hit Auswahl“) wichtig. Da nicht jedes aktiv getestete Molekül das Potential für Weiterentwicklungen hat, ist diese Auswahl von besonderer Bedeutung, um ein Scheitern in den späteren Phasen, was mit sehr hohen Entwicklungskosten einhergeht, zu vermeiden.[37] Insbesondere aber in den nachfolgenden Phasen, in denen zunächst eine Weiterentwicklung dieser Moleküle zu Leitstrukturen („Leitstruktur-Generierung“) mit nachfolgender multidimensionaler Optimierung der Moleküleigenschaften („Leitstruktur-Optimierung“) stattfindet, ist die genaue Aufklärung und Kenntnis von SARs von entscheidender Bedeutung für den Erfolg eines Projektes und die angestrebte Überführung in die präklinische Entwicklung.[38] Des Weiteren ist SAR-Information aber auch im sogenannten „De-novo-Design“, wo eine komplett neue Molekülstruktur basierend auf Kristall-Struktur-Information (strukturbasiert) oder bekannten bioaktiven Molekülen (ligandbasiert) inkrementell aufgebaut bzw. entworfen wird, sehr wertvoll.[39]

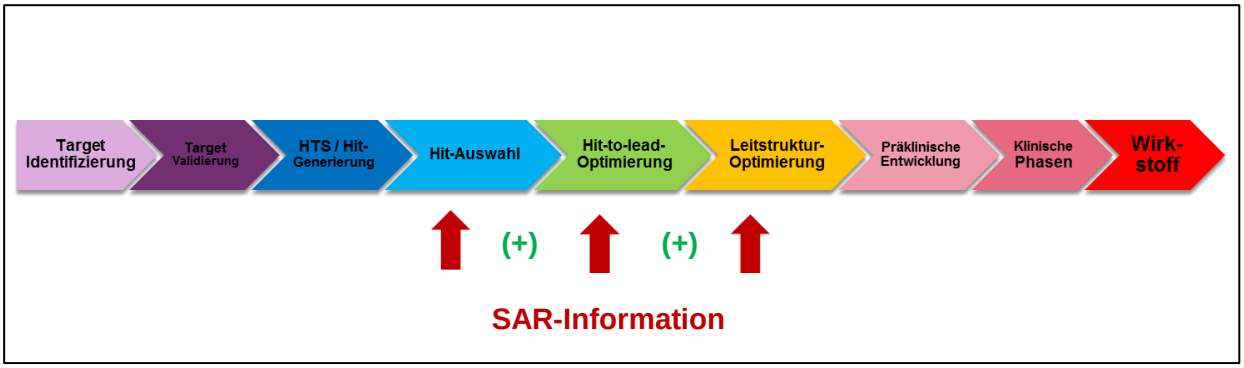


Abbildung 2.1: Die wichtigsten Phasen des Arzneistoffentwicklungs-Prozesses im Fließschema. Die mit roten Pfeilen markierten Entwicklungsstufen profitieren besonders von vorhandener Information über Struktur-Aktivitäts-Beziehungen. Hier können SARs deutlich zur Beschleunigung der Entwicklung beitragen.

2.2. Bioaktivität

2.2.1. Grundlage für biologische Aktivität am Target

Biologische Aktivität ist in erster Linie das Ergebnis der Interaktion eines Moleküls mit seiner makromolekularen Zielstruktur, meist ein Protein. Die Stärke dieser Protein-Ligand-Interaktionen ist weniger an eine bestimmte molekulare Substruktur oder ein bestimmtes Grundgerüst gebunden, sondern abhängig von den physikochemischen Eigenschaften eines Moleküls. Neben elektronischen Eigenschaften spielen räumliche Komplementarität mit der Bindetasche, d.h. sterische Eigenschaften des Moleküls eine große Rolle.[40–41] Zusätzlich hierzu können weitere, z.T. schwer vorhersagbare enthalphische (ΔH) und entropische Effekte (ΔS) Einfluss auf die Änderung der freien Energie (ΔG) bei der Protein-Ligand-Bindung haben.[42] So ist beispielsweise die Auswirkung der Verdrängung von sogenanntem „strukturellen Wasser“ bei der Bindung des Liganden in der Bindetasche sehr schwer abschätzbar.[42–43] Eine nicht-kovalente Interaktion bzw. die Ausbildung eines Protein-Ligand-Komplexes findet nur statt, wenn ΔG negativ ist.[42] Nach Gibbs ergibt sich ΔG aus der Summe des Enthalpie-Terms ΔH und des Entropie-Terms (-TΔS, wobei T die absolute Temperatur in Kelvin darstellt).[42] Schwache Bindungsaffinität (im Folgenden mit „biologischer Aktivität“ gleichgesetzt) eines Liganden an seine Zielstruktur (K_i = 10^{-3} mol l^{-1} = millimolarer Bereich) geht typischerweise mit Werten für die freie Energie um -10 kJ mol^{-1}, extrem hohe Bindungsaffinität (K_i = 10^{-12} mol l^{-1} = picomolarer Bereich) mit Werten für ΔG um -70 kJ mol^{-1} in wässriger Lösung einher.[44]

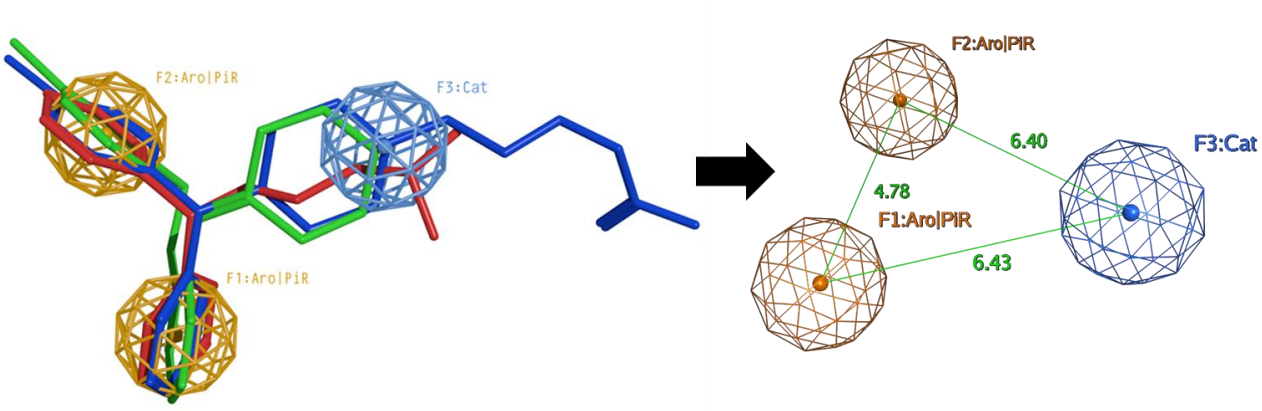


Abbildung 2.2. Links: Superpositionierung der drei strukturell diversen H_1-Antihistaminika Diphenhydramin (rot), Cetirizin (blau) und Desloratadin (grün). Rechts: Daraus abgeleiteter Pharmakophor. Zwei aromatische Ringe (braune Sphären) und ein kationisches Zentrum (blaue Sphäre) in einer bestimmten räumlichen Anordnung (Distanzen in Ångström in Grün) sind für die biologische Aktivität (kompetitiver Antagonismus) am H_1-Rezeptor essentiell.

Das abstrakte Konzept des *Pharmakophors* ermöglicht die Mindest-Voraussetzungen für Bioaktivität basierend auf Interaktions-Eigenschaften eines Moleküls zu definieren. Das bereits 1909 von EHRLICH geprägte Konzept beschreibt die Eigenschaften, die ein Arzneistoff mindestens erfüllen muss, damit er biologische Aktivität an der Zielstruktur

zeigt.[45] Nach IUPAC-Definition von 1998 beschreibt der Pharmakophor eines Moleküls „die räumliche Anordnung sterischer oder elektronischer Eigenschaften, die notwendig sind, um optimale Wechselwirkung mit einer bestimmten biologischen Zielstruktur zu ermöglichen und um eine biologische Antwort auszulösen oder zu blockieren“.[35] Die IUPAC betont dabei, dass ein Pharmakophor ein rein abstraktes Konzept darstelle und nicht durch Nennung bestimmter funktioneller Gruppen (z.B. Sulfonamid) oder Grundgerüste (z.B. Phenothiazin) adäquat beschrieben werden könne.[35] Der Pharmakophor trägt stattdessen den gemeinsamen Interaktions-Eigenschaften (in Form von elektrostatischen, hydrophoben oder H-Brücken-Eigenschaften) mehrerer biologisch aktiver Moleküle Rechnung und stellt den größten strukturellen Nenner dar.[35]

Pharmakophore Eigenschaften beschreiben potentielle Interaktionsmöglichkeiten eines Moleküls. Sie stellen somit einen vielversprechenden Ansatz für die Analyse von SARs dar und bilden daher auch eine wichtige Grundlage der in dieser Arbeit beschriebenen inSARa-Methode. Kapitel 7 gibt daher einen groben Überblick über Möglichkeiten der Definition der wichtigsten pharmakophoren Eigenschaften. Grundlage für die Definition pharmakophorer Eigenschaften, aber auch für die Interpretation von SARs (im Folgenden abgekürzt als „SAR-Interpretation“) im Allgemeinen, sind Protein-Ligand-Interaktionen. Kapitel 6 liefert daher eine kurze Zusammenfassung der wichtigsten Wechselwirkungen nicht-kovalenter Natur.

2.2.2. Gewinnung, Messung und Qualität von Bioaktivitätsdaten

Primär-Screening und Bestätigungs-Testung

Wie in den einleitenden Worten in Kapitel 1 bereits beschrieben, ist das experimentelle Hochdurchsatz-Screening eine Möglichkeit neue potentielle Wirkstoffkandidaten zu finden. Dabei wird innerhalb kürzester Zeit eine große Menge an Bioaktivitätsdaten generiert: Beim klassischen HTS können zwischen 10.000 und 100.000 Moleküle, beim ultra-HTS sogar mehr als 100.000 Moleküle pro Tag getestet werden.[46–47]

Beim Primär-Screening (engl. primary screening) werden i.d.R. viele Moleküle aufgrund des hohen Durchsatzes fälschlicherweise positiv getestet.[37] Bei der Auswertung dieser Daten auch im Hinblick auf vorhandene SARs oder anderweitige Nutzung (z.B. Erstellung von Benchmark-Datensätzen fürs Virtuelle Screening[48]) sollte dies immer berücksichtigt werden. Bei den Falsch-Positiven handelt es sich oftmals um sogenannte „Assay-Artefakte“ aufgrund z.B. von Aggregation[49–54] (meist bei sehr lipophilen Molekülen), optischer Interferenz mit dem Assay System („inner filter effect“) oder unspezifisch-kovalentes Binden[55] beispielsweise verursacht durch reaktive funktionelle Gruppen.[56–59]

Daher werden primäre Hits, d.h. Moleküle mit einer bestimmten Mindestaktivität, zur Bestätigung immer einer Bestätigungs-Testung (engl. confirmatory screening) unterzogen. Falsch negative Moleküle, die in dem Primär-Screening fälschlicherweise inaktiv getestet werden, obwohl sie eigentlich Aktivität an der Zielstruktur zeigen, werden hingegen in der Regel nicht noch mal getestet.[60] Während man bei der Primär-Testung zumeist nur eine einzelne Messung bei einer bestimmten Konzentration durchführt, wird in nachfolgenden Testungen die Aktivität bei mehreren Konzentrationen bestimmt und im Optimalfall ein zum ersten Testsystem orthogonaler Assay verwendet, was die Zuverlässigkeit der resultierenden

Daten deutlich erhöht.[60] Für SAR-Analysen sind generell aufgrund der höheren Datenqualität Daten aus Bestätigungs-Testungen oder Leitstruktur-Optimierungs-Projekten zu bevorzugen.

Assaytypen: Bindungs-Assays und funktionelle Assays

Bei eingesetzten Assays lassen sich grob zwei verschiedene Typen unterscheiden: Bindungs-Assays, die typischerweise an aufgereinigtem Target-Protein („isolierter Target Assay“, Abk. IT) durchgeführt werden, und funktionelle Assays, die meist Zell-basiert (Abk. ZB) sind. Der Bindungs-Assay dient zur Bestimmung der Affinität eines Liganden zum Protein, wobei die Art der Wirkung (z.B. Agonismus oder Antagonismus) im Gegensatz zum funktionellen Assay unklar bleibt. Wichtig ist, dass bei funktionellen Assays im Gegensatz zum Bindungs-Assay immer eine gewisse Unsicherheit bezüglich des Targets besteht und auch allosterisch bindende Moleküle erfasst werden. Eine Zwischenstellung nehmen die membranbasierten funktionellen Assays ein, die z.B. bei G-Protein gekoppelten Rezeptoren (Abk. GPCRs) häufig Anwendung finden. Sie vereinen die Vorteile des IT-Assays (hohe Sicherheit bezüglich des Targets) und des funktionellen Assays (funktionelle Charakterisierung möglich).

Einen detaillierten Überblick liefern HÜSER et al.[61] und MALLENDER et al.[62]. Die wichtigsten Charakteristika sind in der nachfolgenden Tabelle zusammengefasst. Zudem sei auf eine Zusammenstellung von Vor- und Nachteilen von Zell-basierten[63–64] und isolierten Target-Assays[64–65] verwiesen.

Tabelle 2.1. Gegenüberstellung der Merkmale von isolierten Target (IT) und Zell-basierten (ZB) bzw. Bindungs-Assay und funktionellen Assays. Zusammengestellt nach Referenz[61–65].

Assay-Typ	**IT-Assay**		**ZB-Assay**
	Bindungs-Assay	**Funktioneller Assay**	
Prinzip	aufgereinigtes Target-Protein (z.B. kompetitiver Enzym-Substrat-Assay oder Verdrängungsassay mit radioaktiv markiertem Liganden)	mit isoliertem Rezeptor-Protein oder Enzym	meist Zell-basiert (rekombinante Zellen mit Überexpression des Target-Proteins, Reportergen-Assay etc.)
Target-Sicherheit	hoch		gering („off-Targets“ → Falsch-Positiv-Rate↑)
Informations-gehalt	nur isoliertes Target (keine unspezifische Interferenz)		zellulärer Kontext (ggf. Membranpermeation und andere zelluläre Interferenz relevant)
Art der Messung	Bestimmung der Affinität/Bindung des Liganden zum/ an das Protein (Rezeptor oder Enzym), keine funktionelle Charakterisierung	Bestimmung der Art der Wirkung bzw. funktionelle Charakterisierung (z.B. Inhibitor, Aktivator, Antagonist, Agonist)	
Messung	z.B. K_d, K_i, IC_{50}	z.B. IC_{50}, EC_{50}	IC_{50}, EC_{50}, %Hemmung etc.
Orthosterisch/ allosterische Bindung	Erfassung nur von Liganden mit orthosterischer Bindung	Erfassung auch von allosterischen Modulatoren	
Durchsatzrate	High-Throughput möglich (robuster, leichter automatisierbar)		geringerer Durchsatz

In der Phase der Hit-Molekül-Findung sind IT-Assays zu bevorzugen, in der späteren weiteren Charakterisierung eines Moleküls leisten hingegen CB-Assays einen wichtigen Beitrag. Es sei darauf verwiesen, dass nicht für alle Targets (z.B. sogenannte „orphan receptors“ bei den GPCRs ohne bekannten Liganden) IT-Assays möglich sind. Auch sei angemerkt, dass nicht immer eine gute Korrelation zwischen Bindungs-Assay und funktionellem Assay besteht. So zeigen ZHU et al. beispielsweise, dass 9% aller 616 getesteten Moleküle trotz hoher Affinität im Bindungs-Assay nur geringe agonistische Wirkung (Transaktivierung des Pregnan-X-Rezeptors) im Reporter-Gen-Assay zeigen.[66]

Bindungs-Assay: K_i- und IC_{50}-Wert

Sicherheit bezüglich der Zielstruktur ist für SAR-Analysen von hoher Relevanz. In dieser Arbeit wurden daher nur Daten aus Bindungs-Assays verwendet. In den Datenbanken sind hierfür als Bioaktivitäts-Daten zumeist K_i- und IC_{50}-Werte hinterlegt.

Bei reversibler Bindung eines Liganden L an sein Zielprotein P handelt es sich um eine Gleichgewichtsreaktion mit den Geschwindigkeitskonstanten für die Assoziation k_{+1} und die Dissoziation k_{-1}:

$$L + P \underset{k_{-1}}{\overset{k_{+1}}{\rightleftharpoons}} LP \qquad (2.1)$$

Die Bindungsaffinität des Liganden lässt sich über die Gleichgewichts-Dissoziationskonstante K_d charakterisieren, die nach dem Massenwirkungsgesetz den Quotienten aus k_{-1} und k_{+1} darstellt. K_i wird häufig synonym für K_d verwendet, obwohl K_i sich eigentlich auf die Gleichgewichts-Dissoziationskonstante eines Inhibitors bezieht.[67]

Der IC_{50}-Wert gibt die Konzentration an, bei der halbmaximale Enzym-Inhibition auftritt[34] bzw. bei der in kompetitiven Verdrängungsassays (z.B. bei GPCRs) 50% des markierten Referenzliganden verdrängt wird[68]. Der Nachteil von IC_{50}-Werten im Vergleich zu K_i-Werten ist die Abhängigkeit von der verwendeten Substratkonzentration bzw. der Konzentration des Referenzliganden.[67] Im Fall von kompetitiver Enzyminhibition oder kompetitiver Ligand-Rezeptor-Bindung (z.B. bei GPCRs) lässt sich der K_i-Wert aus dem IC_{50}-Wert mittels folgender von CHENG und PRUSOFF[69] entwickelten Gleichung abschätzen, wobei $[L_{Ref}]$ die Substratkonzentration (bei Enzymen) bzw. die Konzentration des Referenzliganden (bei GPCRs) und $K_{d\text{-}Ref}$ die Michaelis-Menten-Konstante (bei Enzymen) bzw. die Gleichgewichts-Dissoziationskonstante des markierten Referenzliganden (bei GPCRs) darstellt[68]:

$$K_i = \frac{IC_{50}}{1 + \frac{[L_{Ref}]}{K_{d-Ref}}} \qquad (2.2)$$

Aktuelle Dimensionen vorhandener Bioaktivitäts-Datenbanken

Aufgrund der in Kapitel 1 beschriebenen Entwicklungen hat sich die Zahl der verfügbaren Bioaktivitätsdaten in den vergangenen Jahren stark erhöht. Nicht nur in den Datenbanken der großen Pharmafirmen sind Millionen von Bioaktivitätsdaten hinterlegt, sondern auch im öffentlich zugänglichen Bereich hat eine große Entwicklung stattgefunden.[70] Letzt genannte sind vor allem für den Bereich der akademischen Forschung von großem Interesse. Die größten, frei zugänglichen Bioaktivitätsdatenbanken stellen PubChem BioAssay[71–72], ChEMBL[73–74] und die BindingDB[75–77] dar.[78] Der folgende Abschnitt soll einen kurzen aktuellen Überblick über die Daten von ausgewählten öffentlichen und kommerziellen Bioaktivitätsdatenbanken geben. Verweise auf eine Vielzahl weiterer Datenbanken geben u.a. OPREA und TROPSHA[79].

PubChem BioAssay enthält über 130 Millionen Bioaktivitätsdaten aus mehr als 717.000 Assays verschiedener Quellen.[72, 80] Im Gegensatz zu ChEMBL und BindingDB handelt es sich hier überwiegend um (Hochdurchsatz-)Screening-Daten (primäre und bestätigende Testungen), was die riesige Menge an verfügbaren Daten erklärt.

In der aktuellsten ChEMBL-Version (Version 17 vom 29.08.2013) sind über 12 Millionen Bioaktivitätsdaten aus fast 735.000 Assays zu mehr als 1,3 Millionen verschiedenen Molekülen und über 9.300 Zielstrukturen vorhanden. Mehr als 7 Millionen dieser Bioaktivitätsdaten stammen dabei aus PubChem BioAssay und mehr als 4 Millionen sind aus der wissenschaftlichen Literatur extrahiert.[81]

Die BindingDB enthält über eine 1 Millionen Bioaktivitätsdaten für mehr als 7.000 Protein-Targets und mehr als 440.000 kleine Moleküle. Der Fokus der BindingDB liegt darauf verfügbare Bindungsaffinitäts-Daten (IC_{50}, K_i, K_d, EC_{50}) über kleine, nicht-kovalent mit Proteinen interagierende Moleküle, in einer großen, öffentlich zugänglichen Datenbank zusammenzustellen. Daten werden zum einen aus der wissenschaftlichen Literatur selber extrahiert und kuriert. Zum anderen werden aber auch ausgewählte Bestätigungs-Assays aus PubChem BioAssay und Einträge aus ChEMBL, für die definierte Target-Informationen vorliegen, in die Datenbank aufgenommen.[76–77] Ein detaillierter Vergleich dieser letztgenannten beiden Datenbanken (ChEMBL und BindingDB) findet sich bei WASSERMANN und BAJORATH.[82]

Eine weitere, öffentlich zugängliche Datenbank von deutlich geringerem Umfang stellt die PDBBind[83–84] dar. Sie stellt eine Sammlung experimentell bestimmter Bindungsaffinitäts-Daten für biomolekulare Komplexe aus der Protein Data Bank (Abk. PDB)[85–86] zur Verfügung. In der ebenfalls frei zugänglichen PDB sind fast 90.000 experimentell ermittelte Protein-Kristallstrukturen hinterlegt, womit sie die größte Quelle für 3-dimensionale Protein-Strukturinformation ist.[86] Die aktuellste Version der PDBBind (Version 2012 vom 30.10.2012) enthält Daten für mehr als 7.100 Protein-Ligand-Komplexe.

Eine weitere bekannte Datenbank stellt die kommerziell von Sunset Molecular Discovery vertriebene WOrld of Molecular BioAcTivity (Abk. WOMBAT) dar.[87–88] In der aktuellsten Version (2013.1 vom 01.02.2013) stellt sie mehr als 335.000 aus der wissenschaftlichen Literatur extrahierte Bioaktivitätsdaten zu fast 2.000 Zielstrukturen bereit.[89]

Vergleichende Analysen verschiedener öffentlicher und kommerzieller Datenbanken u.a. von SOUTHAN et al.[90] (2009) und TIIKKAINEN UND FRANKE[91] (2012) zeigen, dass die einzelnen Datenbanken trotz zu erwartender Überlappungen aufgrund gleicher

Primärquellen doch eine beachtliche Anzahl an einzigartiger Information enthalten. Für SAR-Analysen kann sich somit ein Mehrwert aus der Verwendung von mehreren Datenbanken ergeben. Aber nicht nur von der komplementären Information, sondern auch dem Anteil an überlappender Information können SAR-Analysen profitieren[91], denn sie geben Aufschluss über ein in den vergangenen Jahren immer weiter in den Fokus gerücktes Thema[92–94]: die Qualität der Daten (Datenbankenfehler) und Vertrauenswürdigkeit von bereitgestellten Bioaktivitätsinformationen.

Unsicherheit/Zuverlässigkeit von Bioaktivitätsdaten

Die Zuverlässigkeit bzw. Vergleichsbarkeit von Bioaktivitätsdaten ist ein wichtiger Punkt, den es bei der Zusammenstellung von Daten für SAR-Analysen zu berücksichtigen gilt. Denn selbst wenn man die perfekte Methode (optimale molekulare Repräsentation und Bestimmung von Ähnlichkeit) gefunden hätte, bliebe die Qualität bzw. der Erfolg der SAR-Analysen letztendlich immer durch die zugrunde liegenden Daten limitiert.

Daher haben sich in den letzten Jahren verschiedene Gruppen mit der Analyse der Datenqualität im Allgemeinen[91, 95–97], aber auch im Speziellen mit der Unsicherheit von bereit gestellten Bioaktivitätsdaten[98–100] in verschiedenen kommerziellen, in-house und öffentlich zugänglichen Datenbanken beschäftigt.

Analysen der allgemeinen Fehlerrate in verschieden Datenbanken (z.B. fehlerhafte Einträge in ChEMBL mindestens 4%[94], WOMBAT etwa 8%[87], andere Datenbanken zwischen 0,1 und 3,2%[97]) belegen, wie wichtig es ist, Datenbank-Daten vor der Verwendung nochmal zu standardisieren, auf Konsistenz zu prüfen und ggf. manuell zu kurieren[101]. Die wahre Fehlerrate kann noch höher geschätzt werden, da diese Analysen in der Regel auf der begrenzten Zahl an Mehrfacheinträgen in einer oder mehreren Datenbanken beruhen.[96] Zu den häufigsten Fehlern zählen u.a. Übertragungsfehler bei der Extraktion aus den Primärquellen (v.a. falsche Konvertierung von Einheiten), Strukturfehler wie z.B. falsche Stereochemie oder Konnektivitäten und ungenügende Target-Annotationen.[95–96, 98] Auch das Filtern von Duplikaten ist aufgrund der hohen Zahl an Mehrfacheinträgen für einzelne Moleküle aufgrund automatischer Extraktion aus Primärpublikationen und zitierenden Quellen unerlässlich.[96, 98]

In-house Bioaktivitäts-Daten weisen meist eine geringere experimentelle Unsicherheit auf, die durch die Intra-Laboratoriums-Reproduzierbarkeit eines Assays bestimmt wird (typische Standardabweichungen für pIC_{50}-Messung in dem gleichen Labor liegen etwa bei 0,2 log-Einheiten).[99] Bioaktivitätsdaten aus öffentlichen Datenbanken sind heterogener (verschiedene Labore, Assay-Bedingungen, Assay-Methoden), sodass die Unsicherheit der Daten deutlich größer ist, was verschiedene aktuelle Analysen belegen[98–99] So stellen Analysen der ChEMBL-Datenbank von KRAMER et al. (Version 12) einen mittleren Fehler von 0,44 pK_i-Einheiten und eine Standardabweichung von 0.54 pK_i-Einheiten für heterogene K_i-Daten fest. Analoge Analysen von KALLIOKOSKI von ChEMBL 14 bezüglich der Unsicherheit von heterogenen IC_{50}-Werten ergeben einen mittleren Fehler von 0,55 pIC_{50}-Einheiten und eine Standardabweichung von 0,68 pIC_{50}-Einheiten.[99] Bei SAR-Analysen oder der Beurteilung der Leistungsfähigkeit von in-silico Modellen zur Vorhersage von Bioaktivitäten auf Basis dieser Daten sollten diese Unsicherheiten immer berücksichtigt werden.[98]

K_i-Werte sind im Allgemeinen aufgrund der geringeren Abhängigkeit von den Assay-Bedingungen, (in genannten Analysen wurde dies mit einer etwa 25% geringeren Variabilität von pK_i-Werten im Vergleich zu pIC_{50} belegt[98–99]) und der folglich besseren Vergleichbarkeit von Werten aus verschiedenen Laboratorien (so wie es auch bei den Daten aus den öffentlichen Bioaktivitätsdatenbanken der Fall ist) für SAR-Analysen zu bevorzugen. Dies wurde auch durch STUMPFE und BAJORATH bei der Analyse der BindingDB (März 2011) bestätigt, die ein vermehrtes Auftreten von sprunghaften SARs in Datensätzen, die auf IC_{50}-Werten beruhen (1,04% aller berücksichtigten Molekülpaare im Vergleich zu 0,67% für K_i-Werte), festgestellt haben.[100] IC_{50}-Werte sind jedoch vergleichsweise einfach zu messen und daher das meist genutzte Maß für biologische Aktivität in der Leitstrukturstruktur-Optimierung.[99] Dies ist auch ein Grund für die große Zahl an IC_{50}-Daten im Vergleich zu K_i-Daten in Bioaktivitäts-Datenbanken (in ChEMBL 13 betrug das Verhältnis etwa 2 zu 1[102], in Version 14 bereits etwa 3 zu 1[99]). Würde man IC_{50}-Daten verschiedener Quellen nicht in SAR-Analysen berücksichtigen, würde somit eine riesige Menge an vorhandener und potentiell wichtiger SAR-Information verloren gehen. In dieser Arbeit wurden daher auch heterogene IC_{50}-Daten verwendet, jedoch unter Berücksichtigung der sich für SAR-Interpretation ergebenden Grenzen.

2.3. Molekulare Ähnlichkeit

Zur Beschreibung von molekularer Ähnlichkeit ist es notwendig, sich auf eine Form der Molekülrepräsentation sowie eine Art und Weise, wie die so codierten Moleküle verglichen werden, festzulegen. Eine universell optimale Ähnlichkeitsdefinition zu finden, erweist sich als schwierig. Dies verdeutlicht auch die riesige Anzahl entwickelter bzw. zur Verfügung stehender Möglichkeiten zur Repräsentation von Molekülen bzw. Ähnlichkeits- und. Distanzmaße. In den folgenden beiden Abschnitten 2.3.2 und 2.3.3 werden daher nur die am häufigsten verwendeten kurz beschrieben.

2.3.1. Grundlage von SAR-Analysen: Das Ähnlichkeitsprinzip

Eine der zentralen Annahmen in der medizinischen Chemie beruht auf dem 1990 von JOHNSON und MAGGIORA formulierten Ähnlichkeitsprinzip (Abk. SPP, engl. similar property principle), wonach ähnliche Moleküle auch ähnliche Eigenschaften und somit auch ähnliche biologische Aktivität aufweisen.[103] Das SPP spielt nicht nur bei der Analyse von SARs oder der Modellierung von quantitativen Struktur-Aktivitäts-Beziehungen (vgl. 2.6.1) eine zentrale Rolle. Es bildet u.a. auch die Grundlage für viele auf Ähnlichkeitssuchen basierende Methoden des Virtuellen Screenings[33, 104], das im Arzneistofffindungsprozess eine wichtige Ergänzung zum experimentellen Screening darstellt. Computergestützt wird hierbei auf Grundlage eines bekannten bioaktiven Moleküls in einer virtuellen Substanzbibliothek nach weiteren, zu dieser Vorlage „ähnlichen“ Verbindungen gesucht. Dies geschieht unter der Annahme, dass diese „ähnlichen“ Moleküle dann mit hoher Wahrscheinlichkeit auch Bioaktivität an der entsprechenden Zielstruktur zeigen.[105]

Dass dieses allgemein anerkannte Prinzip nur begrenzte Gültigkeit besitzt, zeigt eine Vielzahl von beschriebenen Beispielen[105–106], wo kleinste Veränderungen an der Molekülstruktur (z.B. das Einführen einer Methylgruppe[107]) zu einem kompletten Verlust der Bioaktivität, aber auch einer massiven Steigerung der Bioaktivität führen können. Dieses gerade in den letzten Jahren vermehrt untersuchte und viel diskutierte Phänomen wird als „sprunghafte SARs“ (Abk. AC, engl. activity cliff) bezeichnet. Details hierzu finden sich in Kapitel 2.5.1. Um das SPP aufrecht zu erhalten, ist es entscheidend molekulare Ähnlichkeit in einer sinnvollen Weise zu erfassen.

2.3.2. Möglichkeiten molekularer Repräsentation

Die Repräsentation von Molekülen ist einer der kritischsten Parameter[108], wenn man auf der Suche nach SAR-Mustern in Datensätzen ist. Es gibt verschiedene Möglichkeiten Moleküle darzustellen. Je nach Komplexität des Modells und des Abstraktionsgrades wird es jedoch immer eine mehr oder weniger starke Vereinfachung der Realität und die adäquate Darstellung immer eine Frage des Kontext bzw. der betrachteten Zielstruktur sein.

Deskriptoren

Eine der einfachsten Formen zur Beschreibung von Molekülen sind Deskriptoren. Hierbei werden Moleküleigenschaften mit Hilfe von numerischen Werten codiert. Ein häufiges Einteilungskriterium ist die Anzahl an Dimensionen, die zur Berechnung des entsprechenden Deskriptors notwendig ist. Während sich *1D-Deskriptoren* von der Summenformel ableiten lassen (z.B. Molekulare Masse, Anzahl eines bestimmten Atomtyps), ist für die Berechnung von *2D-Deskriptoren* die Kenntnis des molekularen Graphen bzw. der Konnektivitäten (2-dimensionale Information) zwischen den Atomen notwendig (z.B. logP, Anzahl an Ringen). *3D-Deskriptoren* (z.B. molekulare Oberfläche, Volumen) hingegen sind von der geometrischen Anordnung der Atome im Raum bzw. der (bioaktiven) Konformation abhängig. Eine ausführliche Zusammenstellung findet sich bei TODESCHINI und CONSONNI.[109]

Fingerprints

Eine besondere, komplexere Form von Deskriptoren stellen die Fingerprints bzw. molekularen Fingerabdrücke (Abk. FP) dar. Hier wird das Molekül als Vektor mit Binärzahlen oder Vektor mit Ganzzahlen repräsentiert, der die An- oder Abwesenheit bzw. die Anzahl des Auftretens einer bestimmten Eigenschaft oder eines bestimmten Strukturmerkmals codiert. Abgesehen davon, dass ebenfalls eine Einteilung in 2D- und 3D-FPs möglich ist, lassen sich folgenden Hauptklassen unterscheiden[110]:

1.) *Wörterbuch-basierte* Fingerprints, wie z.B. die MACCS Keys[111–112], zeichnen sich dadurch aus, dass sie eine vordefinierte Zahl an Substrukturen oder physikalisch-chemischen Eigenschaften repräsentieren.

2.) *Pfad-basierte* Fingerprints, wie z.B. Daylight FP[113] oder Open Babel‘s FP2[114–115], hingegen codieren systematisch Pfade zwischen den einzelnen Atomen auf Grundlage von Atomkonnektivitäten im Molekül.

3.) *Zirkuläre* Fingerprints, wie z.B. der Extended Connectivity Fingerprint[116] (Abk. ECFP), codieren die chemische Umgebung innerhalb eines vordefinierten Radius eines jeden Atoms.

4.) *Pharmakophor*-basierte Fingerabdrücke repräsentieren Pharmakophore auf Grundlage von Tripletts[117] oder Quadrupletts[118] von pharmakophoren Eigenschaften und zugehörigen Distanzen.

Von diesen klassischen Fingerprints, die nur auf Liganden-Information beruhen, sind weitere Typen wie z.B. *Protein-Ligand-Interaktions-Fingerprints*, bei denen zusätzlich Protein-Information berücksichtigt wird, deutlich zu unterscheiden. Als Beispiel ist der von DENG et al. publizierten SIFt (Structural Interaction Fingerprint) zu nennen, wo basierend auf der Information von Protein-Ligand-Komplexen das Interaktionsmuster zwischen dem Liganden und den Aminosäuren der Bindetasche codiert wird.[119]

Repräsentation als Graph (2D- und 3D-Struktur)

Die dem medizinischen Chemiker vertrauteste und daher Standard-Darstellungsweise in der Chemie ist der statische *2-dimensionale molekulare Graph*. Hierbei werden Atome durch Knoten und Bindungen durch Kanten des Graphs repräsentiert (vgl. Kapitel 3). Methoden, die auf 2D-Strukturen basieren sind bei gleicher Leistungsfähigkeit gegenüber 3D-Methoden nicht nur aufgrund des geringeren Rechenaufwandes, sondern auch weil sie für den Chemiker intuitiver erscheinen und leichter verständlich sind, zu bevorzugen.

Ausgehend von dem 2-dimensionalen molekularen Graphen lassen sich kraftfeld- oder wissensbasiert Modelle der *3-dimensionalen Struktur* (Konformerenensembles) bzw. nach Energieminimierung die energetisch günstigste Konformation erzeugen.[120] Hierbei muss es sich nicht notwendigerweise um die bioaktive Konformation handeln.[121] Da molekulare Interaktionen 3-dimensionaler Natur sind bzw. die elektronischen und sterischen Eigenschaften eines Moleküls von der räumlichen Position der Atome im Raum abhängig sind, würde man für Methoden, die auf 3-dimensionalen Molekülstrukturen beruhen, gegenüber 2D-Methoden einen deutlichen Mehrwert in Bezug auf SAR-Analyse (oder im Virtuellen Screening) erwarten. Aufgrund der konformationellen Flexibiliät und des damit einhergehenden erhöhten Rechenaufwandes, sowie der Unsicherheit bezüglich der bioaktiven Konformation kann für 3D-Methoden oftmals keine Überlegenheit gegenüber 2D-Methoden festgestellt werden[121–122].

Darstellung als linearer Textausdruck (SMILES und InCHI)

Die von WEININGER entwickelte *SMILES* Notation (Simplified Molecular Input Line Entry System) dient der kompakten, linearen Repräsentation von Molekülen als leicht lesbaren Textausdruck und kann u.a. dazu genutzt werden, um Ähnlichkeitssuchen zu vereinfachen.[123–124] Durch Anwendung eines Kanonisierungs-Algorithmus auf einen SMILES-Ausdruck kann das entsprechende Molekül in einer einzigartigen, eindeutigen Form repräsentiert werden, was u.a. bei dem Auffinden von Duplikaten oder Substruktursuchen in Datenbanken genutzt werden kann.[124]

Eine weitere Form der linearen Molekül-Repräsentation stellt der von der International Union of Pure and Applied Chemistry und dem NIST entwickelte, frei verfügbare *InCHI* (IUPAC International Chemical Identifier) dar.[125] Hierbei handelt es ebenfalls um einen Textausdruck, der ein Molekül in verschiedenen Ebenen, die Informationen über Konnektivitäten, Ladung, Stereochemie, Isotopie und Tautomerie enthalten können, eindeutig beschreibt.[126] Die Anwendung ist mit den kanonischen SMILES-Strings vergleichbar. Der Vorteil des InCHIs besteht darin, dass auch das Erkennen von tautomeren Formen beim Filtern von Duplikaten möglich ist und dass mit unterschiedlichen Programmen erzeugte InCHIs aufgrund des eindeutigen Algorithmus im Gegensatz zu SMILES problemlos portierbar sind.[127]

Abstraktere Moleküldarstellung (RGs und 3D-Pharmakophore)

Darüber hinaus besteht die Möglichkeit Moleküle zu abstrahieren (z.B. durch Betrachten von pharmakophoren Eigenschaften anstelle von einzelnen Atomen oder Substrukturen), was den Vorteil bietet noch allgemeinere SARs ableiten zu können.

Auf 2-dimensionaler Ebene bietet das Konzept des *reduzierten Graphen* (Abk. RG) die Möglichkeit der abstrakteren Molekülrepräsentation unter Erhalt der Topologie. Da dieses Konzept die Grundlage für die in dieser Arbeit entwickelte inSARa-Methode darstellt, wird in einem eigenen Abschnitt (vgl. Kapitel 5) nochmals detailliert auf reduzierte Graphen eingegangen.

Eine noch stärkere Form der Abstraktion bietet das bereits in Abschnitt 2.2 vorgestellte Konzept des *Pharmakophors*. Hier wird die Topologie nicht mehr berücksichtigt. Es ist nur noch die Anordnung der pharmakophoren Eigenschaften im 3-dimensionalen Raum entscheidend. Ein Nachteil von 3D-Pharmakophoren im Vergleich zu den 2D-RGs ist der größere Rechenaufwand und die Abhängigkeit von der konformationellen Flexibilität. Pharmakophore ermöglichen aufgrund des höheren Abstraktionslevels unter Umständen Gemeinsamkeiten in strukturell sehr diversen Datensätzen zu erkennen.

2.3.3. Quantifizierung von molekularer Ähnlichkeit

Globale Ähnlichkeitsmaße

WILLETT et al. beschreiben eine Vielzahl von Maßen zur Messung von Ähnlichkeit bzw. Distanzen im chemischen Raum.[104] Während die Euklidische Distanz eines der am häufigsten verwendeten Distanzmaße darstellt, ist der der Tanimoto Koeffizient (Abk. Tc, Synonym: Jaccard Koeffient) die am häufigsten verwendete Ähnlichkeitsmaß.[120] Sie werden als „globale Maße“ bezeichnet, da sie keine lokalen Ähnlichkeiten identifizieren.[128]

Die binäre Form des Tanimoto Koeffizient berechnet sich vereinfacht wie folgt[104]:

$$\mathrm{Tc}\,(\mathrm{A},\mathrm{B}) = \frac{\mathrm{N_{AB}}}{\mathrm{N_A} + \mathrm{N_B} + \mathrm{N_{AB}}} \qquad (2.3)$$

Hat man zwei Moleküle A und B, so stellt N_{AB} die Anzahl der Bits dar, die in dem Vektor beider Moleküle gesetzt sind (Wert = „1“). N_A bzw. N_B hingegen ist die jeweilige Anzahl an Bits, die nur in dem Vektor von Molekül A bzw. B gesetzt ist. Der binäre Tc kann Werte zwischen 0 und 1 annehmen, wobei ein Wert von 0 anzeigt, dass keine Ähnlichkeit (bezogen auf den verwendeten Fingerprint) zwischen den beiden verglichenen Molekülen besteht, also keine gemeinsamen Bits vorhanden sind. Ein Wert von 1 bedeutet, dass sich die Moleküle sehr ähnlich sind. Sie müssen aber nicht unbedingt identisch sein, sondern beide Moleküle werden durch den gleichen Fingerprint repräsentiert.

Die Ähnlichkeit zwischen Vektoren mit ganzen Zahlen wird mit der allgemeinen Form des Tc berechnet[104]:

$$\mathrm{Tc}\,(\mathrm{A},\mathrm{B}) = \frac{\sum_{i=1}^{i=n}(x_{iA} \cdot x_{iB})}{\sum_{i=1}^{i=n}(x_{iA})^2 + \sum_{i=1}^{i=n}(x_{iB})^2 - \sum_{i=1}^{i=n}(x_{iA} \cdot x_{iB})} \qquad (2.4)$$

Hierbei werden Molekül A und B durch den Vektor x der Länge n repräsentiert, wobei an Position i jeweils der Wert x_i angenommen wird. Der Tc kann in dieser Form Werte zwischen -0.333 (keine Ähnlichkeit) und +1 (maximale Ähnlichkeit) annehmen.

Lokale Ähnlichkeitsmaße

Um lokale Gemeinsamkeiten zwischen Molekülen zu quantifizieren, ist es immer notwendig die Moleküle aufeinander abzubilden bzw. zu superpositionieren.[120] Diese Methoden sind daher im Vergleich zu den oben beschriebenen Deskriptor- bzw. Fingerprint-basierten Ähnlichkeitsmaßen deutlich rechenaufwändiger, hingegen meist auch besser interpretierbarer. Zudem können auch Ähnlichkeiten zwischen Molekülen, die sich in der Größe deutlich unterscheiden, erkannt werden.[129]

Sehr anschaulich lässt sich die Ähnlichkeit von Molekülen beispielsweise über die maximal gemeinsame Substruktur (Abk. MCS, vgl. Kapitel 4) vergleichen. Je größer die Substruktur ist, die die zu vergleichenden Moleküle gemeinsam haben, umso größer ist die Ähnlichkeit. Basierend auf der Anzahl an Atomen oder Bindungen im MCS können zum Beispiel zum Tc analoge Koeffizienten zur Quantifizierung der Ähnlichkeit verwendet werden. Ein Beispiel hierfür ist der nachfolgend aufgeführte, auf RAYMOND et al. zurückgehende RASCAL Score[130]:

$$\text{RASCAL Score}\,(\mathrm{A},\mathrm{B}) = \frac{(|\mathrm{V}(\mathrm{MCS})| + |(\mathrm{E}(\mathrm{MCS})|)^2}{(|\mathrm{V}(\mathrm{A})| + |(\mathrm{E}(\mathrm{A})|) \cdot\ (|\mathrm{V}(\mathrm{B})| + |(\mathrm{E}(\mathrm{B})|)} \qquad (2.5)$$

|V(X)| ist die Anzahl an Atomen (engl. vertex) der maximal gemeinsamen Substruktur (X=MCS), von Molekül A (X=A) oder Molekül B (X=B). |E(X)| ist entsprechend die Anzahl an Bindungen (engl. edge) in X. Der RASCAL Score kann Werte von 0 (keine Ähnlichkeit) und 1 (maximale Ähnlichkeit) annehmen.[130] Verweise auf weitere Graphen-basierte Koeffizienten finden sich bei RAYMOND et al.[131–133].

2.4. Charakterisierung von SARs

2.4.1. Modell der Aktivitäts-Landschaft

Als Aktivitäts-Landschaft definieren WASSERMANN et al. jede Form der Darstellung, die die Analyse von struktureller Ähnlichkeit und Bioaktivitäts-Unterschieden für Moleküle mit biologischer Aktivität an derselben Zielstruktur vereint.[108]

Die Natur der SAR-Landschaft ist sehr komplex und stark abhängig von der verwendeten molekularen Repräsentation, dem Ähnlichkeitsmaß und dem untersuchten chemischen Raum.[108, 134] Während man früher davon ausging, dass die Aktivitäts-Landschaft dem Bild von sanft geschwungenen Hügeln nahe kommt (MAGGIORA benutzt hierfür als anschaulichen Vergleich die Grasebene von Kansas), musste man in den vergangenen Jahren immer häufiger feststellen, dass es oftmals eher einer zerklüfteten Landschaft ähnelt (nach MAGGIORA vergleichbar mit der des Bryce Canyon in Utah).[134] Dieses wird auf das häufige Auftreten von sprunghaften SARs zurückgeführt.[134]

2.4.2. Quantitative SAR-Charakterisierung

Das SAR-Verhalten eines Datensatzes lässt sich nicht nur qualitativ, sondern auch quantitativ charakterisieren. Hierfür wurden zwei unterschiedliche Maßzahlen entwickelt, die beide auf Fingerprint-basierter Ähnlichkeit beruhen. Zu beachten ist hierbei, dass der zur Molekülrepräsentation verwendete Fingerprint einen großen Einfluss auf die berechneten SAR-Charakteristika haben kann.

Der von PELTASON und BAJORATH entwickelte *SAR-Index* (Abk.: SARI)[135] stellt eine Funktion dar, die sowohl zur Charakterisierung des SAR-Typs eines ganzen Datensatzes (global), als auch einzelner Cluster (lokal) geeignet ist. Der SARI setzt sich aus zwei verschiedenen Komponenten zusammen, dem Kontinuitäts- und der Diskontinuitäts-Score. Der *Kontinuitäts-Score* misst die bioaktivitätsgewichtete strukturelle Diversität aller Datensatzmoleküle. Strukturell diverse Moleküle mit hoher Bioaktivität und kleinen Bioaktivitätsunterschieden werden von dieser Maßzahl hoch bewertet.

$$\mathrm{Kont_{raw}} = \frac{\sum_{\text{alle Ligandenpaare (i,j)}} (w_{i,j} \cdot \mathrm{Tc(i,j)})}{\sum_{\text{alle Ligandenpaare i,j}} w_{i,j}} \qquad (2.6)$$

$$w_{i,j} = \frac{\mathrm{Pot(i)} \cdot \mathrm{Pot(j)}}{1 + |\mathrm{Pot(i)} - \mathrm{Pot(j)}|} \qquad (2.7)$$

Tc(i,j) ist die Tanimoto-Ähnlichkeit für das Molekülpaar i,j des Datensatzes (z.B. unter Verwendung von MACCS Keys[135–136]). Pot(i) bzw. Pot(j) ist die Bioaktivität (angegeben als pK_i oder pIC_{50}) des Moleküls i bzw. j.

Der *Diskontinuitäts-Score* hingegen misst die durchschnittliche Bioaktivitäts-Differenz für Molekülpaare mit hoher struktureller Ähnlichkeit. Bei dieser Maßzahl werden strukturell ähnliche Moleküle mit hohen Bioaktivitätsunterschieden hochbewertet. Um sprunghafte SARs zu erfassen wird eine Mindest-Ähnlichkeitsschwelle definiert und eine minimale Bioaktivitätsdifferenz von einer Größenordnung vorausgesetzt, damit ein Molekülpaar im Diskontinuitäts-Score berücksichtigt wird.

$$\mathrm{Disk_{raw}} = \mathrm{mean}_{\text{alle Ligandenpaare (i,j) mit Tc(i,j)>0.6 \& |Pot(i)− Pot(j)|>1}} \left(|\mathrm{Pot(i)} - \mathrm{Pot(j)}| \cdot \mathrm{Tc(i,j)}\right) \qquad (2.8)$$

Je nach Ausprägung des kontinuierlichen oder diskontinuierlichen Anteils kann der SARI nach Reskalierung des Roh-Kontinuitäts- und -Diskontinuitäts-Scores $Kont/Disk_{raw}$ Werte zwischen 0 (maximale Diskontinuität) und 1 (maximale Kontinuität) annehmen. Der SARI wurde ursprünglich unter Verwendung des MACCS-FP entwickelt, ist aber auf jeden anderen Fingerprint anwendbar.[108, 135] Jedoch ist hierbei zu beachten, dass die für die Umwandlung der Rohwerte ($Kont_{raw}$ und $Disk_{raw}$) in normierte Werte ($Kont_{norm}$ und $Disk_{norm}$) verwendeten Z-Scores (z.B. bei der Berechnung des SAR-Index mittels SARANEA[136]) auf Referenzdatensätzen beruhen, die mit MACCS Keys codiert wurden.[135]

$$\mathrm{SARI} = \frac{1}{2} \cdot (\mathrm{Kont_{norm}} + (1 - \mathrm{Disk_{norm}})) \qquad (2.9)$$

Der von GUHA und VAN DRIE entwickelte *Structure-Activity Landscape Index* (Abk.: SALI)[137] hingegen erfasst nur den diskontinuierlichen Bereich der SAR-Landschaft, also den Teil, wo ähnliche Moleküle große Unterschiede in der Bioaktivität aufweisen. Der SALI ist somit insbesondere zum Auffinden von sprunghaften SARs (vgl. Kapitel 2.5.1) geeignet. SALI wird für jedes Molekülpaar i und j berechnet und setzt die Differenz der Bioaktivitäten A_i und A_j zur paarweisen FP-basierten Tanimoto-Ähnlichkeit Tc ins Verhältnis. Hohe SALI-Werte sind ein Indikator für das Vorhandensein von sprunghaften SARs.

$$\mathrm{SALI_{i,j}} = \frac{|\mathrm{Pot(i)} - \mathrm{Pot(j)}|}{1 - \mathrm{Tc(i,j)}} \qquad (2.10)$$

2.4.3. Einteilung von SAR-Typen

Nach PELTASON und BAJORATH[135] lassen sich wie nachfolgend zusammengefasst drei verschiedene SAR-Typen unterscheiden, die sich mit dem SAR-Index bestimmen lassen.

Der *kontinuierliche* SAR-Typ ist durch hohe SARI-Werte (nahe 1) gekennzeichnet. Er resultiert aus einem hohen Kontinuitäts- und einem niedrigen Diskontinuitäts-Score. Charakteristisch hierfür ist, dass (dem SPP folgend) graduell strukturelle Veränderungen in moderaten Veränderungen der Bioaktivität resultieren. Kontinuierliche SARs sind eine wichtige Grundlage für die Anwendbarkeit von QSAR-Modellen und entsprechende Bioaktivitätsvorhersagen.[108]

Kleine SARI-Werte hingegen resultieren aus einem hohen Diskontinuitäts- und einem kleinen Kontinuitäts-Score und sind charakteristisch für den *diskontinuierlichen* SAR-Typ. Diskontinuierliche SARs sind dadurch gekennzeichnet, dass kleine strukturelle Veränderungen dramatische Auswirkungen auf die Bioaktivität haben. Diese diskontinuierlichen Regionen sind interessante Punkte für Leitstruktur-Optimierung und oftmals sehr reich an SAR Information. Diskontinuierliche SAR-Bereiche schränken den Arbeitsbereich von QSAR-Modellen ein und sind ein häufiger Grund für die schlechte Vorhersagekraft entsprechender Modelle.[134] Ein weiteres Merkmal dieser Regionen ist das häufige Auftreten von sprunghaften SARs.[134]

Der *heterogene* SAR-Typ ist durch mittlere SARI-Werte (etwa 0.5) charakterisiert. Hierbei handelt es sich um eine „variable Aktivitäts-Landschaft“, die eine Kombination von kontinuierlichen und diskontinuierlichen SAR Regionen darstellt. Nach PELTASON und BAJORATH lassen sich zusätzlich hierbei zwei Subtypen unterscheiden.[135] Während bei dem „*heterogen-relaxed*” Subtyp ein hoher Kontinuitäts- und Diskontinuitäts-Score vorliegt, liegen bei dem “*heterogen-constrained*” Typ umgekehrte Verhältnisse vor, also ein kleiner Kontinuitäts- sowie Diskontinuitäts-Score.

2.5. Arten von SAR-Information

2.5.1. Konzept der sprunghaften SARs („activity cliffs“)

Ein wichtiges Phänomen im Zusammenhang mit der SAR-Analyse ist das bereits unter 2.3.1 erwähnte Auftreten von *sprunghaften SARs*, wo kleine strukturelle Veränderungen zu großen Bioaktivitätsdifferenzen führen. Für ein Paar strukturell ähnlicher Moleküle, die große Unterschied in der biologischen Aktivität aufweisen prägte MAGGIORA den Begriff „*activity cliff*“ (dt. sprunghafte SARs, Abk. AC).[134] Dieses Verhalten steht eigentlich im Widerspruch zum SPP (vgl. 2.3.1), weswegen früher angenommen wurde, dass sprunghafte SARs eine eher selten auftretende Erscheinung seien. Zahlreiche Analysen aus der Gruppe von BAJORATH[138–140] aus den letzten Jahren belegen jedoch die These MAGGIORAs, dass sprunghafte SARs einen festen, häufig vorkommenden Bestandteil der SAR-Landschaft darstellen[134]. Einen Überblick über verschiedenste Aspekte und weiterführende Konzepte (z.B. „activity ridges“[141]) im Zusammenhang mit ACs erhält man in aktuellen Reviews[142–143].

ACs sind Segen und Fluch zugleich. Sprunghafte SARs stellen aufgrund ihrer schweren Vorhersagbarkeit (in letzter Zeit wird verstärkt versucht zuverlässige Ansätze zur Vorhersage von ACs zu entwickeln[144–146]) für QSAR-Modelle ein großes Problem dar (vgl. JOHNSON[147]). Sie können jedoch dem medizinischen Chemiker auch wertvolle Hinweise auf Merkmale geben, die entscheidend für die Interaktion eines Liganden mit seiner Zielstruktur sind. Allein das Wissen über das Vorhandensein von ACs ist jedoch eher von geringem Nutzen. Erst ihre Interpretation unter Berücksichtigung des nachbarschaftlichen strukturellen Kontextes liefert wertvolle SAR-Information, die für das Design neuer Moleküle (v.a. in der Leitstruktur-Optimierung) von großem Nutzen sein kann.[142]

Ein großes Problem bei der Identifizierung von ACs ist die Unterscheidung von „echten“ sprunghaften SARs und solchen, die aufgrund von mangelhafter molekularer Repräsentation des entsprechenden chemischen Raums vorgetäuscht werden[148]. Generell stellt die starke Abhängigkeit der Topologie der SAR Landschaft von den verwendeten Deskriptoren bzw. der molekularen Repräsentation und in geringem Maße auch vom Ähnlichkeitsmaß ein großes Problem für die Beurteilung von ACs dar.[134, 149] Es erschwert eine einheitliche, klare AC-Definition zu finden. Aufgrund ihrer größeren Invarianz bezüglich dieser Problematik sind daher sogenannte „*Consensus Activity Cliffs*“ von besonderem Interesse für tiefergehende SAR Analysen. Von einem Consensus AC spricht man, wenn trotz der Variation der verwendeten Deskriptoren oder der Ähnlichkeitsmetrik ein strukturell ähnliches Molekülpaar immer wieder große Unterschiede bezüglich der Bioaktivität aufweist.[150]

Aufgrund des oben genannten Einflusses werden ACs immer in Abhängigkeit von den verwendeten Strukturdeskriptoren definiert. Als sinnvolle Definitionsgrundlage haben sich v.a. der ECFP4 und die MACCS Keys jeweils in Kombination mit dem Tanimoto Koeffizienten erwiesen, da für diese FPs empirisch ermittelte Tc-Schwellenwerte bekannt sind, unter deren Verwendung zumeist eine strukturelle Ähnlichkeit zwischen den verglichenen Molekülen mit dem Auge erkennbar ist und somit chemisch interpretierbare ACs resultieren.[142] Die Verwendung von MMPs (vgl. Abschnitt 2.6.5), also klar-definierter Substrukturbeziehungen, stellt darüber hinaus eine empfehlenswerte Alternative dar.[151] STUMPFE und BAJORATH definieren ACs wie nachfolgend beschrieben. Als „großer

Unterschied in der Bioaktivität“ wird eine Bioaktivitäts-Differenz von mindestens 2 Größenordnungen angesehen und die notwendige „hohe strukturelle Ähnlichkeit“ wird auf Grundlage der nachfolgend aufgeführten Kriterien definiert[140]:

- Für MACCS Keys: Das Molekülpaar muss eine Tc-Ähnlichkeit ≥ 0.85 aufweisen.
- Für ECFP4: Das Molekülpaar muss eine Tc-Ähnlichkeit ≥ 0.55 aufweisen.
- Für MMPs: Der Größen-Unterschied der ausgetauschten Fragmente darf nicht mehr als 8 Nicht-Wasserstoff-Atome betragen und die Maximal-Größe eines der ausgetauschten Fragmente darf 13 Nicht-Wasserstoff-Atome nicht überschreiten. Diese Definition wird zurzeit als der Goldstandard aufgrund der Unabhängigkeit von der Repräsentation und einfachen Interpretierbarkeit angesehen.[143]

ACs können aber nicht nur wie bisher betrachtet ligandbasiert, sondern auch strukturbasiert auf der Grundlage von Protein-Ligand-Komplexen identifiziert werden. Dies ist zum Beispiel in der von SEEBECK und RAREY entwickelten ISAC-Methode („Identification of Structure-based Activity Cliffs“) realisiert, die auf der Verwendung einer Kombination aus dem SALI-Konzept (vgl. 2.4.2) und Protein-Ligand-Interaktions-FPs (vgl. 2.3.2) beruht. Die mittels ISAC gewonnene AC-Information kann u.a. zur Visualisierung von Schlüsselinteraktionen in der Bindetasche durch Hervorheben von Protein-Atomen, die AC-Hotspots darstellen, genutzt werden und so eine wichtige Hilfe beim strukturbasierten Design von neuen Wirkstoff-Molekülen darstellen.[152]

2.5.2. Bioisosterie

Für die Leitstruktur-Optimierung stellt die Methode des „bioisosteren Austausches“ eine Möglichkeit dar, die physikochemischen (z.B. Löslichkeit), toxikologischen oder ADME-Eigenschaften, sowie die synthetische Zugänglichkeit oder aber Selektivität eines Leitstruktur-Kandidaten unter Erhalt der bereits optimierten biologischen Aktivität zu verbessern.[153–154] Einen umfassenden Überblick über das Thema gibt BROWN in einem Buch.[155] Des Weiteren sei auf gute Übersichtsarbeiten verwiesen.[153–154]

Streng definiert sind *klassische Isostere* Moleküle oder Substrukturen, die die gleiche Anzahl an Atomen und die gleiche Anzahl an Valenzelektronen aufweisen.[153] Eine breitere Definition beschränkt sich auf ähnliche physikochemische Eigenschaften.[153] Isostere stellen nicht notwendigerweise Bioisostere dar.[154]

Nach IUPAC-Definition sind *Bioisostere (= nicht-klassische Isostere)* Moleküle, die sich durch den Austausch einer (physikochemisch oder topologisch) ähnlichen Atomgruppe unterscheiden, aber ähnliche biologische Eigenschaften aufweisen.[35] Der Erhalt der biologischen Aktivität trotz geringer lokaler struktureller Ähnlichkeit ist charakteristisch für bioisostere Moleküle. Typischerweise weisen sie ähnliche sterische oder pharmakophore Eigenschaften auf.[153]

Klassische Beispiele für bioiostere Gruppen stellt z.B. der Austausch einer Carbonsäure gegen einen Tetrazolring dar (geringe strukturelle Ähnlichkeit, jedoch beides negativ ionisierbare Gruppen mit ähnlichen sterischen und elektronischen Eigenschaften), sowie der

Austausch eines Phenylringes gegen einen Thiophen-Ring. Es ist jedoch zu betonen, dass Bioisosterie schwer verallgemeinbar ist und stark von der strukturellen Umgebung abhängig ist.[153] Für eine Zusammenstellung einer Vielzahl weiterer Beispiele sei auf MEANWELL[154] und BROWN[155], bei dem auch verschiedenste Methoden und Konzepte zur Identifizierung bioisosterer Gruppen vorgestellt werden, verwiesen.

2.5.3. „SAR Hotspot“

Der Begriff des „SAR Hotspots“ beschreibt das Phänomen, dass strukturell ähnliche Moleküle, die sich z.B. nur in einer Substitutionsstelle unterscheiden, eine hohe Variabilität in der Bioaktivität zeigen.[156–158] Oftmals können an diesen Stellen hoher SAR-Diskontinuität wichtige molekulare Merkmale identifiziert werden, die entscheidenden Einfluss auf die biologische Aktivität an der Zielstruktur haben. SAR Hotspots können dadurch wichtige Information für die SAR-Interpretation liefern.

Von sprunghaften SARs sind SAR Hotspots dadurch abzugrenzen, dass es sich hierbei nicht um ein einzelnes Molekül handelt, das sich von seinen strukturell sehr ähnlichen nächsten Nachbarn durch eine deutliche Bioaktivitätsdifferenz unterscheidet. Bei einem SAR Hotspot ist die Aktivitätslandschaft im Allgemeinen sehr heterogen und eine Vielzahl an ähnlichen Molekülen zeigt deutliche Diversität bezüglich der Bioaktivität. Eine weitere Ursache für das Auftreten von SAR Hotspots kann jedoch auch eine ungenügende molekulare Repräsentation sein. In diesem Fall wären sie eher artifizieller Natur und der resultierende Informationsgewinn eher gering.

2.6. Methoden der SAR-Analyse

Einen umfangreichen Überblick über entwickelte Methoden zur Analyse, Visualisierung und Charakterisierung von Struktur-Aktivitäts-Beziehungen finden sich in einer Reihe von Reviews.[108, 159–162] Im Folgenden werden die wichtigsten Entwicklungen in diesem Bereich kurz zusammengefasst und das Prinzip, sowie Vor- und Nachteile einzelner Methoden hervorgehoben.

2.6.1. QSAR

HANSCH und FUJITA[163] bzw. FREE und WILSON[164] legten bereits 1964 mit ihren grundlegenden Arbeiten den Grundstein für die klassische Analyse der quantitativen Struktur-Aktivitäts-Beziehungen (Abk. QSAR). Seitdem stellt sie eine etablierte Methode des computergestützten Arzneistoff-Entwicklungsprozesses dar und findet v.a. in der Leitstrukturoptimierung ihre Anwendung. QSAR kann als mathematische Modellierung des funktionellen Zusammenhanges Y = f(X) zwischen mit Deskriptoren codierten, molekularen Eigenschaften (X) und gemessener biologischer Aktivität (Y) beschrieben werden. Ziel ist es eine mathematische Gleichung aufzustellen, die eine Vorhersage der Bioaktivität (oder im Bereich der QSPR-Analyse auch physikochemischer, toxikologischer oder pharmakokinetischer Eigenschaften) neuer, unter Umständen nur virtueller Moleküle anhand der entsprechenden Deskriptoren ermöglicht. So können Kosten und Zeit für die aufwändige Synthese von Molekülen gespart werden. Einen Überblick über die Geschichte[165], das Prinzip[166] und die Anwendung und sich ergebende Probleme[167] gibt KUBINYI in verschiedenen Reviews.

Die nachfolgend beschriebenen Methoden und auch die in dieser Arbeit entwickelte inSARa Methode sind deutlich von der klassischen QSAR abzugrenzen, wo die Entwicklung eines auf gute Vorhersagen optimierten Modelles im Vordergrund steht. Bei den nachfolgend genannten Methoden, wie auch bei inSARa, ist jedoch das Primärziel die Identifizierung bzw. qualitative Charakterisierung der vorhandenen SAR-Muster in großen Datenmengen. Sie sind analytisch, deskriptiver anstelle von prädiktiver Natur und daher eher in den Bereich des Data-Mining einzuordnen.

2.6.2. R-Gruppen-basierte Analyse SAR-Analyse

Klassische R-Gruppen-Tabellen

Traditionell werden SARs mit Hilfe sogenannter „R-Gruppen-Analysen“ interpretiert. Auch heute findet diese Form der Analyse noch vielfach (v.a. in der Phase der Leitstrukturoptimierung) Anwendung. Die Moleküle werden hierbei in ein konstantes, gemeinsames Grundgerüst und entsprechend variierende Substituenten (die „R-Gruppen“) unterteilt. Die Bioaktivität wird dann in Abhängigkeit von verschiedenen Substituenten-Kombinationen in Tabellenform dargestellt. Aus diesen „R-Gruppen-Tabellen“ werden dann vom medizinischen Chemiker SARs manuell abgeleitet und so Ideen für Strukturvorschläge für potentielle neue Wirkstoffmoleküle generiert. Der Erfolg dieser Form der SAR-Analyse ist sehr stark von der Erfahrung und Intuition des medizinischen Chemikers abhängig. Ein weiterer Nachteil ist zudem, dass dieses Verfahren nur auf sehr kleine Serien von strukturell analogen Verbindungen (etwa zwischen 10 und 100 Moleküle) angewendet werden kann.[159]

(Erweiterte) SAR Maps

Eine Weiterentwicklung dieser klassischen „R-Gruppen-Tabellen“ stellen die von AGRAFIOTIS et al. entwickelten „SAR Maps“ (engl. „SAR maps“)[168] dar. Sie vereinfachen die Analyse dadurch, dass man damit paarweise verschiedene R-Gruppenkombinationen in Form einer Matrix, deren Felder je nach betrachteter Eigenschaft farblich annotiert sind, interaktiv untersuchen kann. Die „erweiterten SAR Maps“ (engl. „enhanced SAR maps“)[169] verbessern die Grundvariante dadurch, dass neben zusätzlichen Annotationsmöglichkeiten nun mehrere Eigenschaften (z.B. Aktivitäten an verschiedenen Zielstrukturen) simultan in einem Matrixfeld visualisiert werden können. So können z.B. im Vergleich zu den ursprünglichen SAR Karten schnell selektivitätsbestimmende R-Gruppen-Kombinationen in Analogserien identifiziert werden. (Erweiterte) SAR Maps sind maximal für die Analyse von wenigen hundert Molekülen (80-200) geeignet.[168–169]

2.6.3. Scaffold-basierte Analysen

Scaffold-Definition und „Scaffold-Hopping“

Der Begriff des molekularen „Scaffolds“ ist ein weitverbreitetes Konzept in der medizinischen Chemie zur Charakterisierung der chemischen Grundstruktur und zur Einteilung von Molekülen bzw. zur Beschreibung von molekularer Diversität. Eine klare, einheitliche Definition ist schwierig und konnte bisher nicht gefunden werden. So finden sich in der Literatur viele verschiedene Konzepte zur Bestimmung und Beschreibung, wodurch der objektive Vergleich verschiedener Scaffold-basierter Analysen erschwert wird.[170]

Grundlegende Arbeit haben Bemis und Murcko 1996 mit dem vielzitierten Konzepts des sogenannten „Bemis-Murcko Scaffolds“(Abk. BMS) geleistet.[171] Hierbei wird das Molekül systematisch in Seitenketten, Linker und Ringsysteme unterteilt, wobei die beiden letztgenannten Einheiten den Scaffold bilden. Die „zyklischen Gerüste“ (Abk. CSK, engl. cyclic skeletons) und „reduzierten zyklischen Gerüste“ (Abk. RCS, engl. reduced cyclic skeletons) stellen eine weitere, abstraktere Scaffold-Definition dar.[171–172] Sie lassen sich von den entsprechenden BMSs ableiten. Während beim BMS die Information über Atomtypen und Bindungsordnungen erhalten bleibt, repräsentiert der CSK nur noch das entsprechende Kohlenstoffgerüst. RCSs stellen CSKs mit Einheits-Ring- und Einheits-Linker-Größe dar. Der Vorteil dieser Scaffold-Definitionen ist, dass sie objektiv und unabhängig vom verwendeten Datensatz sind.[173] Die maximal gemeinsame Substruktur (vgl. Kapitel 4) stellt ein Beispiel für eine weitere oftmals verwendete Scaffold-Definition (z.B. NICOLAOU et al.[174]) dar, die jedoch den Nachteil aufweist, abhängig vom verwendeten Datensatz zu sein.[173]

Eines der primären Ziele des VS ist das Auffinden von Molekülen mit ähnlicher biologischer Aktivität, aber unterschiedlichen Grundgerüsten. Dies wird häufig mit dem Begriff „Scaffold-Hopping“ beschrieben. In den vergangenen Jahren lässt sich eine inflationäre Zunahme des Begriffes beobachten, was die Bedeutung im modernen Arzneistoffentwicklungsprozess unterstreicht.[153] „Scaffold-Hopping“ kann als ein Spezialfall des bioisosteren Austausches (vgl. 2.5.2) angesehen werden mit der Besonderheit, dass es sich bei dem ausgetauschten Molekülfragment um einen zentral gelegenen Molekülteil handelt.[153] SCHNEIDER definierte den Begriff als Identifizierung von isofunktionellen molekularen Strukturen mit signifikant unterschiedlichem Grundgerüst.[175] Scaffold-Hopping ermöglicht so beispielsweise die Umgehung von patentgeschütztem chemischen Raum.[170] Für einen guten Überblick und weitere Detail sei auf gute Übersichtsartikel von LANGDON et al.[153], BROWN und JACOBY[176] und BÖHM et al.[170] sowie ein aktuelles Buch[177] verwiesen.

Scaffold-Analyse

Aufgrund der oben dargestellten Bedeutung des Scaffold-Konzeptes wird es häufig auch zur SAR-Analyse von Datensätzen verwendet. Einen Überblick über verschiedenste Ansätze geben HU et al.[178] sowie SCHUFFENHAUER und VARIN[179].

Der von SCHUFFENHAUER et al. entwickelte hierarchische Baum aus Ringsystemen namens Scaffold Tree[180] stellt die Grundlage vieler weiterer Ansätze dar. Hierbei werden alle Datensatz-Moleküle ausgehend vom BMS iterativ, regelbasiert immer weiter zu kleineren

Ringsystemen abgebaut bis am Ende nur noch ein einzelner Ring überbleibt. Die verschiedenen Ringsysteme bzw. Scaffolds werden in Form einer Baumstruktur angeordnet, wobei die einzelnen Ringe jeweils die Wurzel bzw. die niedrigste Hierarchieebene darstellen und die jeweiligen Superstruktur-Scaffolds die jeweils höhere Hierarchieebene bilden. Ein ähnlicher hierarchischer Scaffold-basierter Ansatz stellt das von WILKENS et al. entwickelte HierS[181], sowie der Ansatz von CHO und SUN[182], der auf einer hierarchischen Organisation maximal gemeinsamer Grundgerüste beruht. Eine Weiterentwicklung des Scaffold Tree stellen die Scaffold Networks[183] dar.

In Verbindung mit Bioaktivitätsdaten wurde der Scaffold Tree Ansatz schon für zahlreiche SAR-Analysen verwendet (z.B. RENNER et al.[184]). Weitere Beispiele sind der von WETZEL et al. entwickelte Scaffold Hunter[185], der Scaffold Explorer von AGRAFIOTIS und WIENER[186], der eine Kombination aus Scaffold Tree und SAR Maps darstellt oder aber der der SAReport in MOE[187–188], als Kombination aus Scaffold-basierter und R-Gruppen-Analyse angesehen werden kann. Das Besondere ist, dass ähnliche Scaffolds bezüglich ihres Substitutionsmusters verglichen werden können. Der von GUPTA-OSTERMANN et al. entwickelte Layered Skeleton−Scaffold Organization (Abk. LASSO) Graph“[189] stellt einen weiteren Ansatz zur SAR-Analyse basierend auf in einer hierarchischen Graphenstruktur angeordneten Scaffolds dar. Im Gegensatz zu den vorgenannten Analysen wird hierbei nicht der BMS und daraus abgeleitete kleinere Ringsysteme, sondern der CSK-Scaffold verwendet.

Fazit

Das Besondere bei den meisten der vorgestellten Ansätze ist die Verwendung einer hierarchischen Graphen- oder Netzwerk-Struktur, die die Analyse intuitiv gestaltet. Durch Scaffold-basierte Analysen können Ringsysteme, die bevorzugt in bioaktiven Molekülen vorkommen, identifiziert werden. Da jedoch die Substituenten bzw. das Substitutionsmuster der Grundgerüste häufig entscheidend für die Bioaktivität ist, ist eine Analyse nur basierend auf dem Scaffold zum Ableiten von SARs in Datensätzen häufig ungenügend. Daher stellen Ansätze wie der Scaffold Explorer oder der SAReport, die zusätzlich Elemente der klassischen R-Gruppen-Analyse beinhalten, wichtige Weiterentwicklungen dar. Zur Analyse von großen Datensätzen aus mehreren Hundert oder Tausend Molekülen ist der SAReport jedoch nicht geeignet.

2.6.4. Fingerprint-basierte Ansätze

Eine große Anzahl an Methoden zur deskriptiven SAR-Analyse basiert auf Fingerprint-Ähnlichkeit. Die wichtigsten Ansätze werden im Folgenden kurz vorgestellt.

SAS Maps

Die von SHANMUGASUNDARAM und MAGGIORA entwickelten *„Structure−Activity Similarity Maps“* (Abk. SAS Map) stellen eine der ersten Methoden zur Charakterisierung der Aktivitätslandschaft in Form einer 2-dimensionalen Karte dar.[190] In SAS Maps wird die paarweise Fingerprint- und Bioaktivitäts-Ähnlichkeit aller Moleküle gegeneinander aufgetragen. Je nach Lage eines Datenpunkes in der Karte kann unterschiedliche SAR-Information basierend auf Molekülpaaren wie „Scaffold-Hopping“, sprunghafte SARs oder (dis-)kontinuierliches Verhalten identifiziert werden. Aufgrund des paarweisen Vergleiches und der somit exponentiell steigenden Zahl an Datenpunkten bei zunehmender Datensatzgröße, sind SAS Maps jedoch nicht für die intuitive Analyse großer Datensätze geeignet.

Graph- und Netzwerk-basierte Ansätze

Die nachfolgend beschriebenen Ansätze haben gemeinsam, dass sie im Gegensatz zu den SAS Maps auf Graphen- oder Netzwerk-Repräsentation basieren. Diese Form der Darstellung findet in den letzten Jahren zunehmende Anwendung für die Visualisierung von Informationen nicht nur im Bereich der SAR-Analyse, sondern auch diversen anderen Bereichen der Arzneimittelforschung, weil so komplexe Zusammenhänge schnell erfasst und eine große Menge von Daten auf intuitive Weise dargestellt werden können.[191–192]

SALI-Graph

Um einen besseren Überblick über das diskontinuierliche SAR-Verhalten in Datensätzen zu erhalten, können die für einen Datensatz berechneten SALI-Werte (vgl. Abschnitt 2.4.3) auch in Form eines Graphen darstellt werden. In diesen *„SALI-Graphen“* wird jedes Molekül als Knoten repräsentiert. Zwei Knoten werden über eine gerichtete Kante (vom schwächer zum höher aktiven Molekül) verknüpft, wenn der SALI-Wert für dieses Molekülpaar einen benutzerdefinierten Schwellenwert überschreitet. Kanten zeigen das Vorhandensein von sprunghaften SARs an. Auch eine Darstellung der SALI-Werte als Heatmap kann zum besseren Erfassen von diskontinuierlichen SAR-Charakteristika beitragen.[137]

NSGs

Der von WAWER et al. entwickelte „*Network-like Similarity Graph*“ (Abk. NSG) Ansatz beruht auf dem Prinzip des Fingerprint-basierten Ähnlichkeitsnetzwerkes, in denen man einen Überblick über die Bioaktivitätsveränderungen strukturell ähnlicher Molekülen erhält.[193] In

den NSGs wird jedes Molekül des Datensatzes als einzelner Knoten dargestellt. Zwei Knoten werden durch eine ungerichtete Kante miteinander verbunden, sofern die zuvor berechnete Fingerprint-Ähnlichkeit zwischen den beiden Molekülen, die durch diese Knoten jeweils repräsentiert werden, einen bestimmten Ähnlichkeits-Schwellenwert überschreitet. Dieser Schwellenwert wird in Abhängigkeit vom verwendeten FP-Typ festgelegt (z.B. MACCS Keys: 0.65-0.75[193–194]; ECFP4: 0.4-0.55[195]). Des Weiteren werden die Knoten entsprechend ihrer Bioaktivität eingefärbt. Die Knotengröße wird in Abhängigkeit von dem lokalen Diskontinuitäts-Score, der für jedes Molekül berechnet wird, skaliert. Ein hoher Diskontinuitäts-Wert, repräsentiert durch einen großen Knoten, zeigt dabei an, dass das betrachtete Molekül sich deutlich bezüglich der Bioaktivität von seinen strukturellen Nachbarn im jeweiligen betrachteten chemischen Raum unterscheidet. NSGs wurden für die SAR-Analyse von Datensätzen bestehend aus ein paar Hundert Molekülen erfolgreich verwendet.[193] Da die NSGs mit steigender Datensatzgröße aufgrund der nicht-hierarchischen Struktur sehr komplex und unübersichtlich werden, ist die visuelle Analyse in sehr großen Datensätzen wenig intuitiv und Auffinden wichtiger SARs gestaltet sich schwierig.[160–161]

Erweiterungen für die Analyse von NSGs

Für diesen Zweck wurden die folgenden drei Erweiterungen, die jeweils Subgraphen des NSG darstellen, entwickelt. Mit ihrer Hilfe ist es möglich systematisch die in den komplexen NSGs verborgene SAR-Information zu extrahieren.

Ein sogenannter *„SAR Pathway“* ist nach WAWER et al. definiert als eine Sequenz von paarweise ähnlichen Molekülen, die steigende biologische Aktivität aufweisen.[194] SAR Pathways werden durch Analyse von Pfaden, also Sequenzen miteinander über Kanten verbundener Knoten, in den entsprechenden NSGs identifiziert. Da SAR Pathways zur Identifizierung von kontinuierlichem SAR-Verhalten entwickelt wurden, werden Pfade nur in Richtung steigender Bioaktivität durchlaufen. Da zwischen einem vordefinierten Start- und Endknoten oftmals eine Vielzahl verschiedener SAR Pathways enumeriert werden kann, ist es notwendig diese mit Hilfe einer Bewertungsfunktion zu priorisieren. Hierfür wird jeder Kante des Pfades entsprechend der strukturellen Ähnlichkeit und der Bioaktivitätsdifferenz mit Hilfe einer Kostenfunktion ein bestimmter Kostenbetrag zugeordnet und diese Kosten dann für alle Kanten des Pfades aufsummiert. Der Pfad mit den geringsten Kosten, der mittels des kürzesten-Pfad-Algorithmus von Dijkstra[196] ermittelt wird, wird bevorzugt. Bestimmt man systematisch zwischen jedem möglichen Start- und Endknoten im Graphen den geringste-Kosten-Pfad, erhält man eine Vielzahl an SAR Pathways, die wiederum bezüglich enthaltener SAR-Information bewertet werden müssen. Diese Bewertungsfunktion weist längeren Pfaden, die eine große Gesamt-Bioaktivitätsdifferenz aufweisen und bei denen der Bioaktivitätsanstieg pro Kante möglichst gering ist, einen hohen Rang zu. Pfade, die mehr als 30% gleiche Knoten enthalten, werden im Anschluss noch gefiltert, um möglichst wenig redundante SAR-Information zu extrahieren.[197]

Ein *„SAR Tree“* ist eine Menge von (hoch bewerteten) SAR Pathways, die den gleichen Startknoten aufweisen, organisiert in einer Baumstruktur. Der gemeinsame Startknoten stellt hierbei die Wurzel des Baumes dar, während die Endknoten aller Pfade die Blätter repräsentieren. SAR Trees eignen sich dazu, dass SAR-Verhalten in der Umgebung eines bestimmten Moleküls von Interesse genauer zu studieren.[197]

Der sogenannte *„Chemical Neighborhood Graph“* (Abk. CNG) stellt eine weitere Möglichkeit dar, die chemische Nachbarschaft (definiert durch einen bestimmten Fingerprint-Ähnlichkeitsschwellenwert) eines bestimmten Moleküls zu untersuchen.[136] Alle chemischen Nachbarn werden hierzu kreisförmig um das entsprechende Referenzmolekül, das den Mittelpunkt darstellt, angeordnet, wobei der jeweilige Radius mit abnehmender struktureller Ähnlichkeit zunimmt. Ähnlich wie bei den SAR Pathways beschrieben, ist es auch hierbei möglich alle CNGs für einen Datensatz berechnen zu lassen und anschließend bezüglich enthaltener SAR-Information mittels einer weiteren Bewertungsfunktion zu priorisieren.[198]

Es konnte gezeigt werden, dass NSGs in Kombination mit diesen Erweiterungen auch für die Analyse von größeren und strukturell heterogeneren Datenmengen verwendet werden können. So konnten WAWER et al. auf diese Weise erfolgreich fünf verschiedene HTS-Datensätze (Bestätigungs-Testungen) aus PubChem bestehend aus mehreren tausend Molekülen (1766 bis 5817) analysieren.[194, 197] WAWER und BAJORATH haben das NSG-Konzept in Kombination mit Similarity-Potency Trees[199] ebenfalls erfolgreich zur Analyse von phänotypischen HTS-Daten eines Antimalaria-Screenings der Firma GlaxoSmithKline[200] bestehend aus über 13500 biologisch aktiven Molekülen, die Vorhersagen zufolge an mehr als 145 verschiedenen mikrobiellen Zielstrukturen angreifen, verwendet.[195]

SARANEA

SARANEA vereint die zuvor vorgestellten Graphen-basierten Ansätze, die in der Gruppe von BAJORATH (NSGs und die NSG-basierte Datenstrukturen) entwickelt wurden. Es handelt sich um ein freiverfügbares, Java-basiertes Programm mit graphischer Benutzeroberfläche zur interaktiven SAR-Analyse. Zusätzlich zu verschiedenen Graphen-Repräsentationen ermöglicht es die globale und lokale Charakterisierung des SAR-Typs eines Datensatzes mittels SARI-Berechnung (vgl. 2.4.2). Zudem kann es auch zur Interpretation von Struktur-Selektivitäts-Beziehungen verwendet werden. Der Input, der vom Benutzer bereitgestellt werden muss, ist auf keinen bestimmten Fingerprint- oder Bioaktivitäts-Typ festgelegt.[136]

SPT

Einen weiteren Ansatz zur Graphen-basierten SAR-Analyse, der auf Fingerprint-Ähnlichkeit beruht, stellen die ebenfalls von WAWER und BAJORATH entwickelten *„Similarity-Potency Trees“* (Abk. SPTs) dar.[199] Wie bei den SAR Trees wird hierbei eine Baum-Struktur zur Repräsentation verwendet und eine lokale SAR-Betrachtung der Nachbarschaft eines Referenzmoleküls, das die Wurzel des Baumes darstellt, ermöglicht. Im Gegensatz zu den SAR Trees, wo die Bioaktivität von der Wurzel ausgehend entweder stetig zu- oder abnimmt und die Blätter den Endpunkt des Bioaktivitäts-Gradienten darstellen, beruhen SPTs auf dem Prinzip des strukturell nächsten Nachbarn und die strukturelle Ähnlichkeit nimmt ausgehend von der Wurzel stetig ab. In den SPTs werden nur die Datensatz-Moleküle als Knoten dargestellt, die eine strukturelle Ähnlichkeit oberhalb eines bestimmten Schwellenwertes bezogen auf das betrachtete Referenzmolekül aufweisen. Diese Moleküle werden dann mit ihrem jeweils nächsten Nachbarn verknüpft. Um ein Baumlayout sicher zu stellen, werden im Fall von mehreren möglichen nächsten Nachbarn weitere Regeln zur Priorisierung eines Nachbarn angewendet. Zur systematischen Analyse eines Datensatzes kann für jedes Molekül der zugehörige SPT berechnet und anschließend mit einer Bewertungsfunktion

bezüglich der enthaltenen SAR-Information priorisiert werden. Hierfür wird für jeden Knoten ein Score berechnet und anschließend diese Scores für alle Knoten aufsummiert. Als besonders reich an SAR-Information (hoher Score) werden Knoten betrachtet, wenn sie möglichst viele strukturelle Nachbarn aufweisen, deren Bioaktivitätswerte stark vom eigenen Wert abweichen. SPTs können sowohl zur Analyse von Datensätzen, die aus Optimierungsprojekten stammen, als auch von heterogeneren HTS-Daten verschiedenster Größe (s.o.) verwendet werden.[195, 199]

CAG

Im Gegensatz zu den bisher vorgestellten Methoden ist der von PELTASON et al. entwickelte „Combinatorial Analog Graph“ (Abk. CAG) Ansatz auf die Analyse von Serien chemisch strukturell analoger Verbindungen fokussiert, was v.a. im Bereich der Leitstruktur-Optimierung von Interesse ist.[156] Das Prinzip des CAG beruht darauf, dass variierende Substitutionsstellen in analogen Verbindungen hierarchisch organisiert werden, sodass Substituenten, die entscheidend die Bioaktivität beeinflussen, leichter identifiziert werden können. Hierfür werden zunächst Analogserien im Datensatz mittels Scaffold-Analyse identifiziert. Alle Moleküle, die das gleiche Grundgerüst aufweisen, werden derselben Serie zugeordnet. Innerhalb jeder Serie wird die maximal gemeinsame Substruktur (vgl. Kapitel 4) bestimmt, auf deren Grundlage die zugehörigen Moleküle in variable R-Gruppen mit einheitlicher Nummerierung und konstantes, gemeinsames Grundgerüst zerlegt werden. Für Molekül-Subgruppen, die sich jeweils an den gleichen Substitutionsstellen unterscheiden, werden jeweils SARI-Diskontinuitäts-Scores berechnet. Anschließend werden diese verschiedenen Subgruppen in einer hierarchischen Graphen-Struktur organisiert. Im Gegensatz zu anderen Ansätzen repräsentieren hier Knoten kein einzelnes Molekül, sondern immer Gruppen von Molekülen, die ein bestimmtes gleiches Merkmal aufweisen. Die Wurzel des Graphen repräsentiert die ganze Molekül-Serie. Auf den folgenden Hierarchie-Ebenen nimmt die Zahl der variierenden Substitutionsstellen immer weiter zu. Knoten werden mit der nächsthöheren Hierarchieebene durch eine Kante verbunden, sofern sie gleiche Substitutionsstellen aufweisen. Um die Interpretation zu vereinfachen werden die einzelnen Knoten zusätzlich gemäß des vorher berechneten Diskontinuitäts-Wertes eingefärbt. Knoten mit hohem Diskontinuitäts-Wert weisen auf SAR Hotspots (vgl. 2.5.3) hin. Fehlendende Substituenten-Kombinationen können aufgrund der hierarchischen Struktur ebenfalls leicht identifiziert werden und geben einen Hinweis auf bisher nicht untersuchte SAR-Regionen, sogenannte „SAR-Löcher“.[156]

mtCAG

WASSERMANN et al. haben den ursprünglichen CAG Ansatz[156] für die Analyse von multi-Target SARs erweitert.[157] So können einfacher selektivitätsbestimmende Merkmale beim Vergleich mehrerer verwandter Zielstrukturen erkannt, aber auch SARs an einem einzelnen Target allgemeiner interpretiert werden. Im sogenannten „multi-target Combinatorial Analog Graph“ (Abk. mtCAG) werden dazu die variablen Substituenten am gemeinsamen Grundgerüst jeweils in Anlehnung an HARPER et al. und die von ihm verwendete RG-Implementierung[201] (vgl. Kapitel 5) als pharmakophore Eigenschaften codiert. Der Diskontinuitäts-Score wird nun nicht mehr basierend auf der Fingerprint-Ähnlichkeit des

ganzen Moleküls berechnet. Stattdessen wird die paarweise Ähnlichkeit in Anlehnung an HARPER et al.[201] mittels einer Kostenmatrix für den gegenseitigen Austausch von Eigenschaften bestimmt, die die Ähnlichkeit von pharmakophoren Merkmalen bewertet. Auf diese Weise lässt sich für das Vorhandensein von pharmakophoren Eigenschaften an einer bestimmten Substitutionsstelle eine bevorzugte Reihenfolge bezüglich bioaktivitäts- oder selektivitätssteigernden Effektes ermitteln, sodass einfacher strukturoptimierende Merkmale identifiziert werden können.

Fazit

Wie an der Vielzahl in diesem Abschnitt beschriebener Ansätze zu sehen ist, stellen auf Fingerprint-Ähnlichkeit beruhende Methoden eine etablierte und häufig verwendete Form der SAR-Analyse dar. Fingerprint-Ähnlichkeit hat den Vorteil, dass sie schnell zu berechnen ist und die entsprechenden Methoden somit auch für die Verarbeitung großer Datenmengen gut geeignet sind. Ein Nachteil dieses Konzeptes ist jedoch, dass es häufig schwierig ist, die molekularen Merkmale zu identifizieren, auf denen die berechnete Ähnlichkeit beruht. Dies ist dadurch begründet, dass die berechneten Werte häufig Ganz-Molekül-Ähnlichkeit reflektieren und weniger lokale Gemeinsamkeiten. Die resultierenden molekularen Ähnlichkeitsbeziehungen sind daher u.U. wenig intuitiv, sodass die SAR-Interpretation sich für den medizinischen Chemiker als schwierig erweist. Ein weiteres Problem ist die große Abhängigkeit von dem für die molekulare Repräsentation verwendeten Fingerprint-Typ.

2.6.5. Substruktur-basierte Ansätze

Intuitivere und besser interpretierbare Ansätze verwenden daher klar-definierte Substruktur-Beziehungen anstelle von berechneten Ähnlichkeitswerten. Zwei vielversprechende und miteinander verwandte Konzepte, die hierfür verwendet werden können, stellen die „maximal gemeinsame Substruktur“ (Abk. MCS, engl. „maximum common substructure“) und das „zusammenpassende Molekülpaar“ (Abk. MMP, engl. „matched molecular pair“) dar.

MMP-Analyse

Ein „zusammenpassendes Molekülpaar“ ist definiert als ein Molekülpaar, das sich nur durch eine kleine, wohldefinierte strukturelle Modifikation an einer bestimmten Stelle im Molekül unterscheidet.[202] Das bedeutet, dass beide Moleküle, die zusammen ein MMP bilden, eine große gemeinsame Grundstruktur aufweisen und die molekulare Transformation, die zur Umwandlung des einen in das andere Molekül notwendig ist, sehr klein und i.d.R. in der Größe limitiert ist. Die MMP-Analyse (Abk. MMPA) ermöglicht eine direkte Verknüpfung von Chemie und gemessener Eigenschaft oder Aktivität und ermöglich dadurch eine klare Interpretierbarkeit der Ergebnisse.[203] Mittels MMPA lassen sich allgemeine quantitative Daumenregeln ableiten, die eine Abschätzung ermöglichen, welche Auswirkung eine bestimmte molekulare Transformation (z.B. Austausch eines Wasserstoff-Atoms gegen eine OH-Gruppe oder Austausch eines Phenyl-Ringes gegen ein Thiophen) auf eine bestimmte Eigenschaft (wie die biologische Aktivität, Wasserlöslichkeit, Plasma-Protein-Bindung) haben könnte.[203] Wichtig ist die Einbeziehung des strukturellen Kontextes in der Umgebung der molekularen Transformation (vgl. PAPADOPATOS et al.[204]), um die MMPA möglichst spezifisch zu machen und die Aussagekraft der abgeleiteten Regeln zu steigern. Durch Vergleich analoger Molekül-Serien lassen sich beispielsweise Vorschläge für den Entwurf neuer Moleküle machen („SAR-Transfer“)[205] bzw. molekulare Transformationen vorhersagen, die eine bestimmte Eigenschaftsänderung bewirken („inverse QSAR“)[206]. WARNER et al. konnten prospektiv eine Überlegenheit der MMPA-Vorhersagen gegenüber einem Random Forest QSAR-Modell zeigen.[206]

Algorithmen zur Bestimmung MMPs lassen sich grob in drei Kategorien einteilen.[207] Limitiert lassen sich MMPs unter Verwendung einer Liste vordefinierter Transformationen (z.B. LEACH et al.[202]) bestimmen. Die MMP-Identifizierung durch die Bestimmung des MCS (z.B. SHERIDAN[208], SHERIDAN et al.[209] oder WARNER et al.[210]) ist mit hohem Rechenaufwand verbunden, hat aber den Vorteil der Identifizierung auch von molekularen Transformationen an mehreren Stellen im Molekül. Besonders effizient sind Algorithmen, die auf Fragmentierung basieren (z.B. HUSSAIN und REA[211]), jedoch bleibt hier trotz Erkennung von nicht-zusammenhängenden gemeinsamen Substrukturen die Einschränkung auf Molekülpaare, die nur an einer Stelle variieren.

Für weitere Vor- und Nachteile dieser Algorithmen, sowie einen umfassenden Überblick über verschiedene Anwendungen des Konzeptes der MMPA im Bereich der (Q)SAR/QSPR-Analyse sei auf gute, aktuelle Übersichtsarbeiten von GRIFFEN et al.[203] und WASSERMANN et al.[207] verwiesen.

BMMSG

Der von WAWER und BAJORATH entwickelte „Bipartite Matching Molecular Series Graph“ (Abk. BMMSG) stellt eine Netzwerk-Struktur basierend auf dem MMP-Konzept dar.[158] Mittels Fragmentierung von azyklischen Einzelbindungen werden zunächst alle MMPs bestimmt. Ausgehend von dieser MMP-Analyse lassen sich dann „zusammenpassende Molekül-Serien“ (Abk. MMS, engl. “matching molecular series”) identifizieren, der jeweils alle Moleküle angehören, die sich nur durch eine Modifikation an einer bestimmten Stelle voneinander unterscheiden. Diese MMS werden dann mit den zugehörigen Molekülen in einer bipartiten Graphen-Struktur wie nachfolgend beschrieben organisiert. In den Netzwerken lassen sich zwei Knotentypen unterscheiden: MMS-Knoten, die die gemeinsame(n) Substruktur(en) repräsentieren und entsprechend der Zahl der notwendigen Fragmentierung farblich annotiert werden, und Knoten, die einzelne Datensatzmoleküle repräsentieren und entsprechend der Bioaktivität eingefärbt werden. Die Molekül-Knoten werden mit den MMS-Knoten durch Kanten verbunden, sofern die zugehörige Struktur eine Substruktur des Moleküls darstellt. Da einzelne Moleküle zu mehreren MMS gehören können, entsteht zunächst eine komplexe Struktur, die anschließend zur Steigerung der Übersichtlichkeit und Interpretierbarkeit prozessiert wird. Hierbei wird redundante, weniger spezifische Information entfernt und bestimmte Knoten (z.B. durch Verwendung von Superknoten) zusammengefasst. Außerdem werden zusätzlich hierarchische Beziehungen zwischen bestimmten Substrukturen in einer separaten Baum-ähnlichen Graphen-Struktur gezeigt. Anhand eines aus 881 Molekülen bestehenden Faktor-Xa-Inhibitoren Datensatz konnte exemplarisch gezeigt werden, dass mittels BMMSG-Analyse wichtige SAR-Trends und -Informationen (wie z.B. SAR Hotspots oder sprunghafte SARs) auch in großen Datensätzen erfolgreich auf der Grundlage von lokalen Substruktur-Beziehungen systematisch identifiziert werden können.[158]

Fazit

MMPAs und BMGSS ermöglichen intuitive, direkt interpretierbare SAR-Analyse. Limitiert sind diese Ansätze jedoch dadurch, dass sie auf Molekülpaare mit nur geringen Unterschieden in der Struktur angewiesen sind. Schon mehrere kleine Variationen in der ansonsten gemeinsamen Substruktur stellen ein Problem für diese Methoden dar, da sie auf exakte Paarungen angewiesen. Sie sind somit nützlich für die SAR-Analyse von Analog- und Parallelserien (anderer Scaffold, gleiches Substitutionsmuster). Für die Analyse strukturell heterogenerer Daten sind sie nicht geeignet.

2.6.6. RG-basierte SAR-Analyse

Wie in Abschnitt 2.3.2 beschrieben stellen reduzierte Graphen eine weitere, abstraktere Möglichkeit der molekularen Repräsentation dar. Dieser Abstraktionsgrad bietet einen besonderen Vorteil für die Anwendung im Bereich der SAR-Analyse. Insbesondere bei strukturell sehr diversen Datensätzen mit vielen verschiedenen Grundgerüsttypen, bei denen die im vorigen Abschnitt beschriebenen intuitiven Substruktur-basierten Verfahren in ihrer Leistungsfähigkeit limitiert sind, sind RG-basierte Ansätze eine vielversprechende Alternative. RGs ermöglichen pharmakophore Gemeinsamkeiten trotz struktureller Unterschiede zu erkennen und somit allgemeinere SAR-Regeln abzuleiten. Dieser Vorteil wird auch bei den im vorigen Abschnitt beschriebenen mtCAGs genutzt, wo die variablen Substituenten als pharmakophore Eigenschaften codiert werden. Bisherige RG-basierte Ansätze zur SAR-Analyse stellen oftmals eine Kombination des RG-Konzeptes mit Techniken des maschinellen Lernens dar.[212–214]

Kombination mit Entscheidungsbäumen

BARKER et al. haben gezeigt, dass als Fingerprints codierte RGs in Kombination mit Entscheidungsbäumen für die Erstellung von SAR-Modellen verwendet werden können.[212] Dies wurde erfolgreich am Beispiel eines 2D-Pharmakophor-Modelles für Angiotensin-II-Rezeptor-Antagonisten demonstriert. Ein Nachteil bei der Verwendung von Fingerprints ist, dass die Information über die Verknüpfung der codierten Knoten-Kanten-Paare bzw. die genaue Molekül-Topologie verloren geht und so z.T. uneindeutige Modelle entstehen können.[215]

Kombination mit evolutionärer Optimierung

BIRCHALL et al. haben RGs in Form von Pseudo-SMARTS (vgl. 10.2.3) mit einem evolutionären Algorithmus für die Erstellung von SAR-Modellen kombiniert.[213] Der entwickelte Algorithmus erstellt RG-Pseudo-SMARTS, die dann als Köder für die Klassifizierung eines Satzes an Trainingsmolekülen, die ebenfalls als RGs codiert sind, verwendet werden. Die erstellten SMARTS werden bezüglich ihrer Fähigkeit bewertet, zwischen aktiven und inaktiven Trainingsmolekülen unterscheiden zu können. Als Gütemaß werden Genauigkeit und Sensitivität verwendet. Mittels genetischer Programmierung werden die SMARTS Köder iterativ optimiert. SMARTS ermöglichen eine größere Flexibilität im Vergleich zu fixen Substruktursuchen mittels SMILES-Repräsentation, sodass eine größere Zahl strukturell unterschiedlicher Moleküle gefunden werden kann. Außerdem ermöglichen sie eine kompakte, dennoch detaillierte SAR-Beschreibung, die für den medizinischen Chemiker direkt interpretierbar ist. Da bei vielen Zielstrukturen eine hohe Diversität bei den Grundgerüsten der biologisch aktive Moleküle festzustellen ist, ist es schwierig die Eigenschaften aller aktiven Moleküle in einem SMARTS-String zu codieren. Aus diesem Grund haben BIRCHALL et al. den ursprünglichen Ansatz erweitert, sodass mehrere SMARTS Köder und folglich verschiedene SAR-Modelle mittels multiobjektiver Optimierung entwickelt werden können.[214] Um möglichst komplementäre Modelle zu generieren, wurde

zusätzlich noch ein Einzigartigkeits-Score zur Bewertung der Redundanz innerhalb der erzeugten Pseudo-SMARTS Köder eingeführt.

Kodierung von bioisosteren Gruppen für RG-basiertes Virtuelles Screening

BIRCHALL et al. haben RGs ebenfalls zur Codierung bioisosterer Gruppen mit dem Ziel einer erhöhten Leistungsfähigkeit von Ähnlichkeits-basierten Virtuellem Screening verwendet.[216] Anhand von Analysen der BIOSTER Datenbank[217], die Paare bioisosterer Moleküle enthält, konnte gezeigt werden, dass RGs in vielen Fällen in der Lage sind bioisostere Gruppen in Form von gleichen Knotentypen zu codieren. Anhand von einer Vielzahl an bioisosteren Paaren, in denen Bioisostere durch eine unterschiedliche Knotentypen oder eine ungleiche Anzahl an Knoten repräsentiert wurden, konnte jedoch auch aufgezeigt werden, dass es schwierig ist eine allgemein gültige RG-Definition zum Erkennen von bioisosterem Austausch zu finden. Eine Überlegenheit dieses Ansatzes im Virtuellen Screening auf Basis der WOMBAT-Datenbank konnte trotz verschiedener Modifikationen im Vergleich zu FCFP4-Fingerprints nicht erzielt werden, obwohl Scaffold-Hopping Potential nachgewiesen werden konnte. Durch die Definition von äquivalenten Knotentypen wird zwar das Erkennen von strukturell diverseren äquivalenten Gruppen ermöglicht, diese Flexibilität und die Abstraktion der RGs führt jedoch auch dazu, dass unter Umständen wenig sinnvolle Beziehungen hergestellt werden.[215] Dies hat sich im Virtuellen Screening in einer verstärkten Anreicherung von inaktiven Molekülen geäußert.

Fazit

Trotz aufgezeigter Grenzen unterstreichen die bisherigen Ansätze und Analysen dennoch das Potential von RGs für die Anwendung in Bereich der SAR-Analyse. Insbesondere durch die Reduzierung des Moleküls auf pharmakophore Eigenschaften, können gemeinsame Eigenschaften, die für Protein-Ligand-Interaktionen wichtig sind, auch in heterogeneren Datensätzen erkannt werden. Entscheidend ist (wie die Analyse von BIRCHALL et al. bestätigt[216]) die Wahl des für SAR-Analysen angemessenen Abstraktionsgrades bzw. die Anpassung der verwendeten RG-Definition.

3. Graphentheoretische Grundlagen

Im Folgenden sind einige graphentheoretische Definitionen, die zum Verständnis der in dieser Arbeit beschriebenen Konzepte benötigt werden, nach DIESTEL[218], TITTMANN[219] und TURAU[220] zusammengefasst.

Graphentheorie stellt ein Teilgebiet der Mathematik dar, das sich mit den Eigenschaften verschiedener Graphen und deren Beziehungen zueinander beschäftigt. Auch Moleküle lassen sich mathematisch als Graphen beschreiben und auf diese Weise miteinander vergleichen. In der Graphentheorie wird das gesamte Molekül als *einfacher, ungerichteter Graph G = (V,E)* betrachtet, wobei die Atome die *Knoten*menge *V={u,v,…}* (engl. vertex) und die Bindungen die *Kanten*menge *E={e,f,…}* (engl. edge) des Graphen darstellen. Ein *einfacher Graph* ist ein Graph, der weder Schlingen (d.h. es gibt eine Kante, die denselben Start- und Endknoten *u* hat) noch Parallelen (d.h. es gibt mehrere Kanten zwischen denselben Knoten *u* und *v*) enthält. Ein *(un)gerichteter Graph* ist ein Graph, der aus einer Menge (un)geordneter Knotenpaare, d.h. (un)gerichteter Kanten, besteht. Eine *gerichtete Kante* wird durch ein geordnetes Knotenpaar, dem Anfangs- und Endknoten, bestimmt. Durch die *Färbung* der Knoten bzw. Kanten können verschiedene Atom- bzw. Bindungstypen unterschieden werden.

Da Graphen oftmals auf verschiedene Weise dargestellt werden können, ist es schwierig zu erkennen, ob es sich um die gleichen Graphen handelt. Um z.B. zu überprüfen, ob es bei Molekülen um Duplikate handelt, wird graphentheoretisch überprüft, ob die beiden zugehörigen Graphen *isomorph* sind. Unter *Graphen-Isomorphismus* zwischen zwei Graphen G_1 und G_2 versteht man, dass die Knoten des einen Graphen auf die Knoten eines anderen bijektiv (d.h. umkehrbar eindeutig) abgebildet werden können unter Berücksichtigung der Gleichartigkeit von verbindenden Kanten und Färbungen der zugehörigen Knoten (quasi G_1 ist gleich G_2). Isomorphe Graphen lassen sich einfacher als Graphen beschreiben, die gleiche Knoten- und Kantenzahl aufweisen und in denen gleiche Knoten mit gleichen Kanten verbunden sind.

Substrukturen von einem Molekül entsprechen in der Graphentheorie Subgraphen. Ein *Subgraph* G_1 von G enthält jeweils eine Untermenge der Knoten- und Kantenmenge von G. Jeder Graph G‘, für den der Graph G einen Subgraphen darstellt, wird wiederum als *Supergraph* von G bezeichnet (analog zur Superstruktur in der Chemie). Ob eine bestimmte Substruktur in einem Molekül enthalten ist (Substruktur-Suche), wird graphentheoretisch überprüft, ob ein Subgraph-Isomorphismus vorliegt. Ein *Subgraph-Isomorphismus* zwischen zwei Graphen ist ein Isomorphismus zwischen dem Graphen G_1 und einem Subgraphen H_1 von H (quasi G_1 ist ein Teil von H bzw. G_1 ist gleich H_1). Das Erkennen eines Subgraph-Isomorphismus ist noch rechenaufwändiger als das Erkennen eines Graphen-Isomorphismus, da initial viel mehr mögliche Startzuordnungen existieren und geprüft werden müssen. Ein *(Knoten-)induzierter Subgraph* besteht aus einer Knotenmenge S eines Graphen G, bei der alle Kanten der Knoten S auch in G enthalten sind. Da es sich um einen Untergraph handelt, stellt S stellt eine Untermenge der Knotenmenge von G dar.

Ein vollständiger, ungerichteter (Sub-)Graph von G wird auch als *Clique* von G bezeichnet. Ein *vollständiger Graph* zeichnet sich dadurch aus, dass jedes Knotenpaar miteinander durch eine Kante verbunden ist. Viele chemoinformatische Probleme lassen sich über die

Bestimmung der maximalen Clique lösen. Ein sehr effizienter Algorithmus geht auf BRON und KERBROSCH[221] zurück.

Ein Graph *G = (V,E)* heißt *zusammenhängend*, wenn es von jedem Knoten *u* zu jedem Knoten *v* des Graphen einen Weg (d.h. einen Kantenzug) gibt. Der zusammenhängende Teilgraph eines nicht-zusammenhängenden Graphen wird als (Zusammenhangs-) Komponente bezeichnet. Ein zusammenhängender Graph besteht aus nur einer Komponente.

Eine Folge durch Kanten miteinander verbundener, adjazenter (d.h. benachbarter) Knoten in einem Graphen wird als *Weg* oder *Pfad* bezeichnet (synomymer Gebrauch beider Begriffe in der Literatur), sofern alle verwendeten Kanten verschieden sind. Ein Weg wird als *geschlossen* bezeichnet, sofern der Anfangs- und Endknoten des Weges gleich ist. Andernfalls handelt es sich um einen *offenen* Weg.

Ein *Baum* ist ein zusammenhängender Graph, der keine Kreise (d.h. keinen geschlossenen Weg) enthält. Für einen Baum mit *n* Knoten gilt immer, dass er *n-1* Kanten besitzt. Ein Graph, dessen Komponenten jeweils Bäume sind, wird als *Wald* bezeichnet. In einem *gerichteten Baum* wird der Startknoten (d.h. der Knoten, von dem ausgehend man jeden anderen Knoten des Baumes erreichen kann) als *Wurzel* bezeichnet. Die terminalen Knoten werden *Blätter* und die übrigen Knoten als *Ast-Knoten genannt*. Sind *u* und *v* Knoten eines gerichteten Baumes und *v* ist von *u* erreichbar, dann wird *u* als *Vorgänger* von *v* und *v* als *Nachfolger* von *u* bezeichnet. Bäume werden häufig zur Darstellung von hierarchischen Strukturen verwendet.

Ein *aufspannender Baum* bzw. *Spannbaum* eines zusammenhängenden, nicht-kreisfreien Graphen G ist ein Subgraph von G, der einen Baum darstellt und die gleiche Knotenmenge wie G aufweist. Analog ist ein *aufspannender Wald* eines nicht zusammenhängenden Graphen G ein Subgraph mit den gleichen Komponenten wie G, der ein Wald ist

Der *minimal aufspannende Baum* oder *Minimale Spannbaum* (Abk. MST, engl. Minimum Spanning Tree) ist der aufspannende Baum eines kantenbewerteten, zusammenhängenden, ungerichteten Graphen, für den kein anderer Spannbaum mit geringeren Kosten existiert. Unter Kosten versteht man hierbei die Summe der Kantenbewertungen des Spannbaums. Zwei häufig verwendete Algorithmen zur Bestimmung des MST sind der Algorithmus von PRIM[222] und der Algorithmus von KRUSKAL[223].

Um nach dem Algorithmus von Kruskal aus einem kantenbewerteten, zusammenhängenden Graphen G mit n Knoten einen MST zu bestimmen, wird wie folgt verfahren (für Details vgl. Abbildung 3.1 und Abbildung 3.2). Ausgangspunkt ist ein kantenfreier Wald B der die gleiche Knotenmenge wie G hat. Als nächstes werden alle Kanten von G nach aufsteigendem Kantengewicht (Bewertung) sortiert. Im Folgenden wird diese Kantenliste in aufsteigender Reihenfolge abgearbeitet. In jedem Schritt wird die Kante mit dem kleinsten Gewicht genommen und in B eingefügt, sofern hierdurch kein geschlossener Weg, also B ein kreisfreier Graph bleibt. Ansonsten wird die Kante verworfen. Der Algorithmus endet, wenn B aus einer Zusammenhangskomponente besteht, also aus *n-1* Kanten besteht.

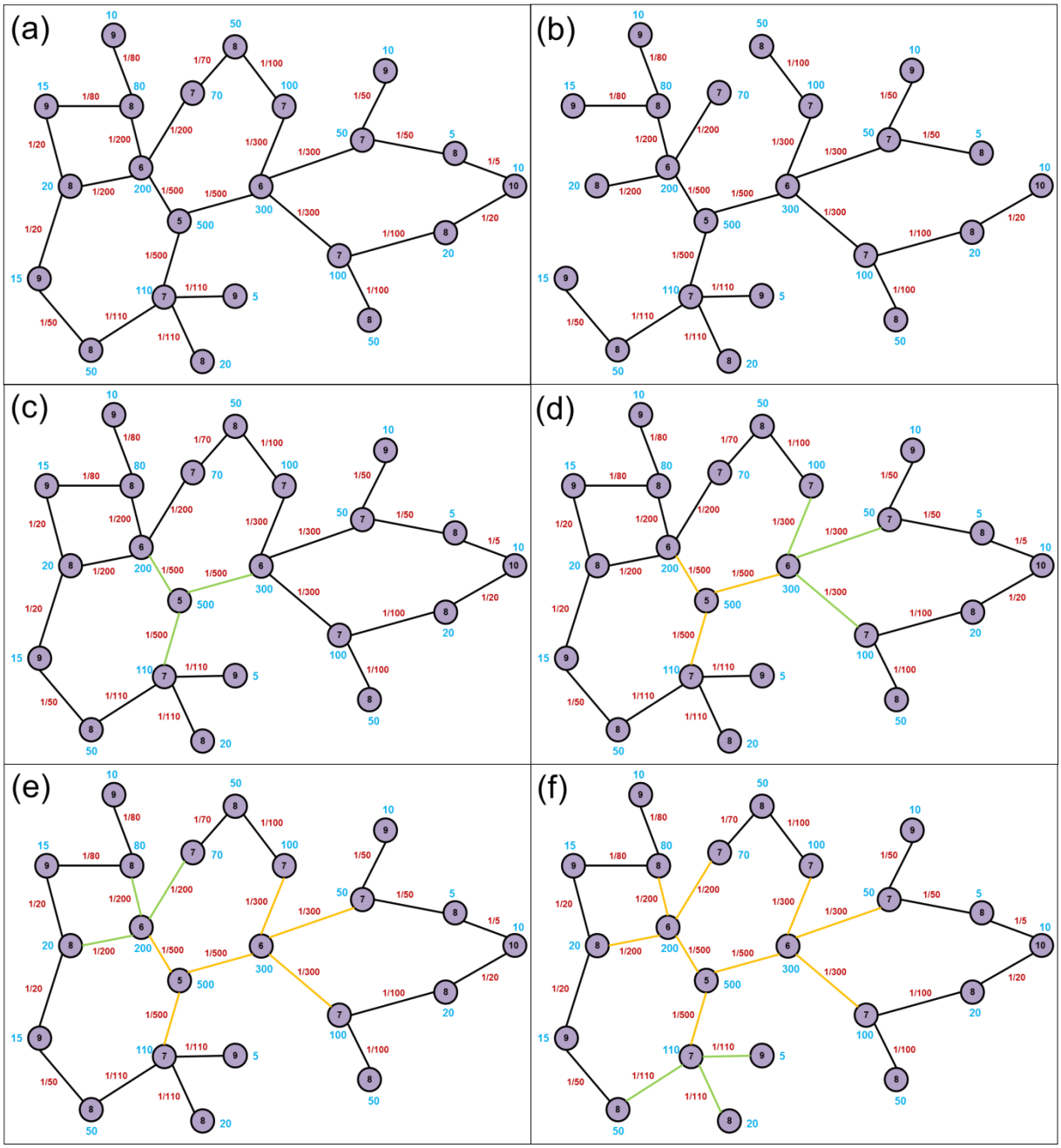


Abbildung 3.1. Prinzip des Algorithmus von Kruskal. Mit dem Verfahren wird der in (b) dargestellte MST des in (a) dargestellten, zusammenhängenden, kantengewichteten (Gewicht = rote Zahl) Graphen G bestimmt (minimales Kantensummengewicht). (c) Beginn des Algorithmus: Der Wald B, der aus denselben Knoten (lila) besteht wie G, die schwarzen Kanten sind nicht Bestandteil von B. Es werden immer die Kanten aus G mit dem kleinsten Gewicht sukzessive wie in (c) bis (f) dargestellt zu B hinzugefügt (grün = neue Kante in B, gelb = bereits vorhandene Kante in B). Fortsetzung in Abbildung 3.2. Die Knotengewichte (blau) und die sich darausergebenden Kantengewichte (rot) werden detailliert in Abbildung 10.6 in Kapitel 10.4 erklärt.

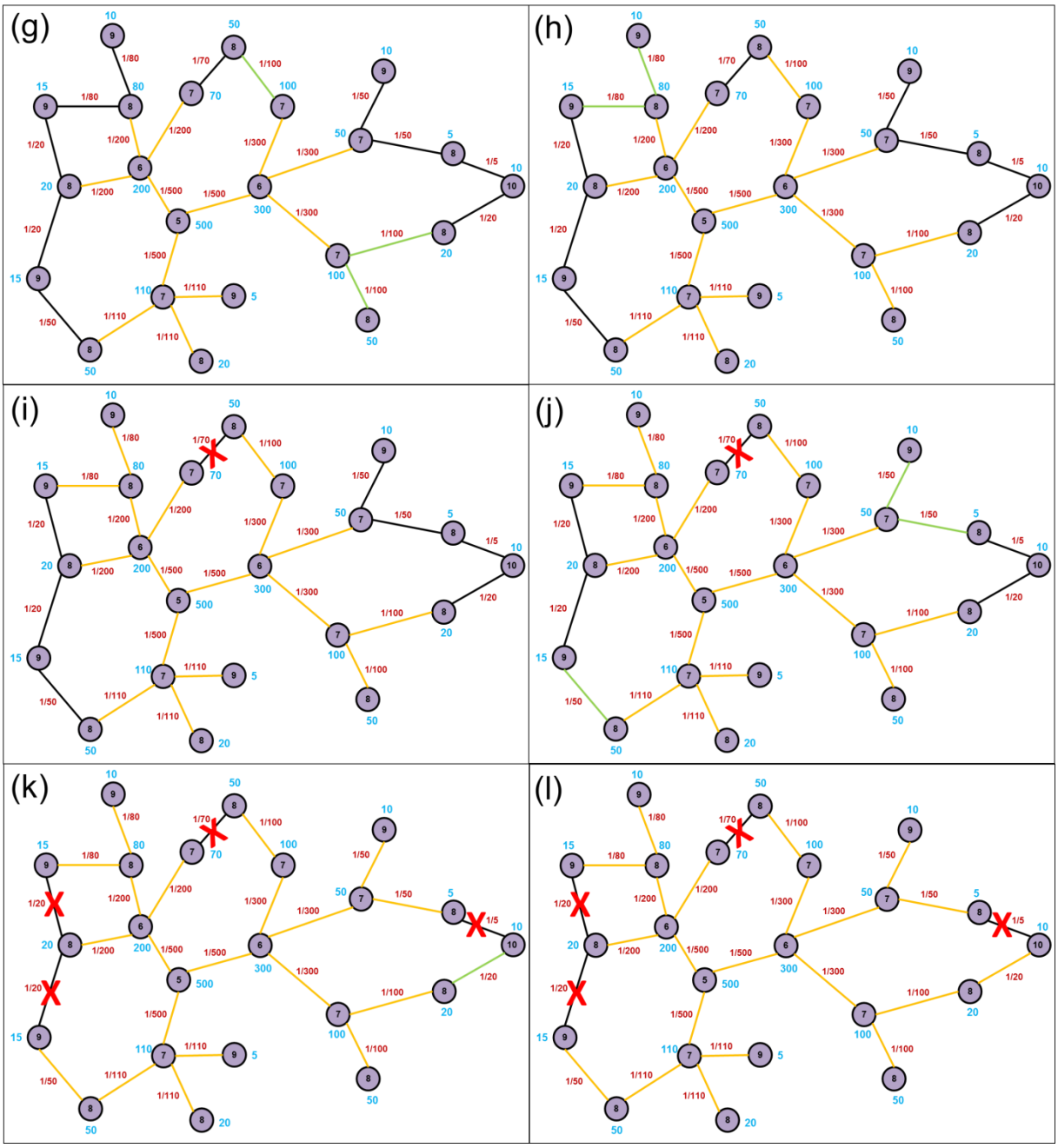


Abbildung 3.2. Prinzip des Kruskal Algorithmus. Fortsetzung aus Abbildung 3.1. In (g) bis (h) werden immer weiter neue Kanten hinzugefügt (grün) zu dem Wald B (lila Knoten, gelbe Kanten) hinzugefügt, sofern kein Kreis entsteht. Durch Hinzufügen der Kante mit dem Gewicht 1/70 in (i) würde diese Bedingung nicht mehr erfüllt sein, daher wird die Kante verworfen (rotes Kreuz). Dieses Prinzip wird solange fortgesetzt (vgl. (j) bis (l)), bis ein zusammenhängender Graph entstanden ist (vgl. (l)). Die Kante mit dem Gewicht 1/5 in (i) braucht nicht mehr berücksichtigt werden, da B (Zahl der Knoten = 22) bereits durch das Hinzufügen der letzten Kante (grün) aus 21 Kanten (= zusammenhängender Graph) besteht.

4. Konzept der maximalen gemeinsamen Substruktur (MCS)

4.1. Definition und Algorithmen

Definitionen

Die „maximale gemeinsame Substruktur“ (Abk. MCS, engl. maximum common substructure) lässt sich allgemein definieren als die größte gemeinsame Substruktur eines Molekülpaares. In der Graphentheorie versteht man unter dem MCS den „maximal gemeinsamen Subgraphen“. Es lassen sich grundsätzlich zwei verschiedene MCS-Typen unterscheiden, der MCIS und der MCES. Die nachfolgenden Definitionen orientieren sich an RAYMOND und WILLETT[132].

Der *maximale gemeinsame induzierte Subgraph* (Abk. MCIS, engl. maximum common induced subgraph) ist definiert als derjenige gemeinsame induzierte Subgraph G_{12} der Graphen G_1 und G_2 mit der größten Anzahl an Knoten (vgl. Abbildung 4.1c). Ein *gemeinsamer induzierter Subgraph* G_{12} zeichnet sich dadurch aus, dass er *isomorph* zu den induzierten Subgraphen von G_1 und G_2 ist.

Zu unterscheiden vom MCIS ist der *maximale gemeinsame Kanten-Subgraph* (Abk. MCES, engl. maximum common edge subgraph). Der MCES ist definiert als derjenige gemeinsame Teilgraph von G_1 und G_2 mit der größten Anzahl an Kanten (vgl. Abbildung 4.1d). Vergleicht man beide MCS-Typen stellt man fest, dass der MCIS wenig intuitiv ist. Unter Berücksichtigung der Bedeutung der Knoten und Kanten in chemischen Graphen, erscheint der MCES zum Erkennen von maximalen molekularen Gemeinsamkeiten zwischen Molekülen chemisch sinnvoller.[132]

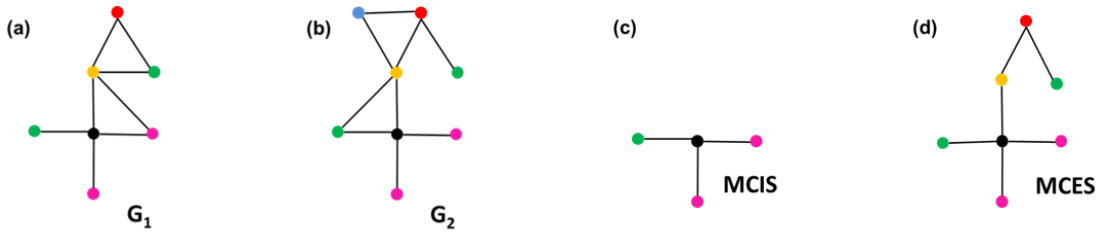


Abbildung 4.1. Zwei Graphen G_1 (a) und G_2 (b), sowie der zugehörige MCIS (c) bzw. der zugehörige MCES (d).

Abgesehen von der Unterscheidung zwischen MCES und MCIS ist es wichtig zwischen *zusammenhängendem* und *nicht-zusammenhängendem* MCS unterscheiden. Im ersten Fall besteht der MCS aus nur einer zusammenhängenden Komponente, d.h. es existiert zwischen jedem Knoten des MCS ein Pfad zu jedem anderen Knoten. Beim nicht-zusammenhängenden Fall hingegen besteht der MCS aus mehreren nicht-zusammenhängenden Komponenten.

Komplexität und Algorithmen

Aufgrund der Komplexität des Subgraph-Isomorphismus, der zur Bestimmung des MCS notwendig ist, handelt es sich um ein sehr rechenintensives Verfahren. Die Berechnung des MCS gehört zu den sogenannten NP-vollständigen Problemen, die für den allgemeinen Fall aufgrund der kombinatorischen Komplexität der notwendigen Vergleiche kein Algorithmus mit polynomieller Zeit-Komplexität verfügbar ist.[132] Betrachtet man beispielweise zwei Moleküle bestehend aus a und b Atomen, so wären kombinatorisch $\frac{a! \cdot b!}{(a-c)! \cdot (b-c)! \cdot c!}$ Atom-Vergleiche notwendig, um alle aus c Atomen bestehende Subgraphen zu bestimmen.[132] Man kann sehen, dass für sehr große Moleküle (d.h. großer Wert für a und/oder b) oder Molekülpaare mit einem großen MCS, also sehr ähnliche Moleküle, (d.h. großer Wert für c) die Berechnungen sehr (zeit-)aufwändig werden.

In der Literatur ist eine Vielzahl verschiedener Algorithmen zur Bestimmung des MCS beschrieben. RAYMOND und WILLETT[132] geben eine gute Übersicht über verfügbare Algorithmen und ihre Anwendungsbereiche. Eine wichtige Unterscheidung bei den Algorithmen ist, ob es sich um ein *näherungsweises* oder *exaktes* Verfahren handelt, das i.d.R. mit einem größeren Rechenaufwand verbunden ist. Aufgrund der Besonderheit chemischer Graphen (i.d.R. wenig komplexe Graphen) im Vergleich zum allgemeinen Graphenfall können zumeist einige Annahmen getroffen werden, die die Berechnungen deutlich vereinfachen, sodass eine Problemlösung in polynomieller Zeit möglich ist. Ein besonders effizienter Algorithmus zur Bestimmung des MCES ist der RASCAL (RApid Similarity CALculation) Algorithmus[130].

4.2. Anwendung

Nicht nur in der Chemoinformatik, sondern auch in vielen anderen Bereichen, wie z.B. der Mustererkennung[224–225] und dem maschinellen Sehen[226–227], findet das intuitive Konzept des maximal gemeinsamen Subgraph erfolgreich Anwendung. In der Chemieinformatik ist das Anwendungsspektrum groß und umfasst z.B. die automatische Auswertung von Spektren (z.B. NMR-^{13}C[228]), die Suche in chemischen Datenbanken[133], das intuitive Gruppieren von Molekülen[229–231], die Clusteranalyse[232] oder aber die SAR- bzw. MMP-Analyse[209–210]. Für weitere Anwendungen im Bereich der Chemie und Biologie sei auf WILLETT[233] verwiesen.

4.3. Vergleich mit Fingerprint-basierter Ähnlichkeit

Der Vorteil von Fingerprints ist, dass sie schnell zu berechnen sind und daher sehr gut für die Analyse großer Datensätze geeignet sind. Jedoch ist häufig schwierig zu verstehen, welche molekularen Eigenschaften für die angenommene Ähnlichkeit verantwortlich sind. Außerdem stellen die berechneten Ähnlichkeitswerte häufig globale Ähnlichkeiten dar, die das ganze Molekül berücksichtigen (vgl. Abschnitt 2.6.4), wodurch unter Umständen wenig intuitive Beziehungen resultieren. Zusätzlich gibt es eine hohe Abhängigkeit von dem Fingerprint-Typ, der für die molekulare Repräsentation verwendet wird. Außerdem kann es

sein, dass wie in Abbildung 4.2 (anhand von zwei FXa-Inhibitoren veranschaulicht) nur eine geringe Ähnlichkeit gefunden wird, obwohl eine große gemeinsame Substruktur vorhanden ist. Die SAR-Interpretation kann somit schwierig für den medizinischen Chemiker sein.

Ähnlichkeit basierend auf dem Konzept des MCS zu definieren ist hingegen intuitiver. Wie in Abschnitt 2.3.3 (im Zusammenhang mit dem RASCAL Score) erwähnt und anhand von Abbildung 4.2 leicht zu erkennen, wird beim MCS-Konzept lokale Ähnlichkeit betont. Es ist leicht nachzuvollziehen, dass eine große gemeinsame Substruktur im Vergleich zur Größe der beiden zu vergleichenden Moleküle eine große Ähnlichkeit anzeigt. Jedoch weist auch das MCS-Konzept einige Nachteile auf. Durch den hohen Rechenaufwand (siehe oben) ist es für die Anwendung auf größere Datensätze ungeeignet, insbesondere, wenn sie aus analogen Molekülen mit einer großen Anzahl an Atomen bestehen.[132] Die Notwendigkeit einer exakten Paarung kann zu sehr kleinen zusammenhängenden MCSs resultieren, wenn beispielsweise Moleküle mit kleinen strukturellen Variationen, aber offensichtlich ähnlichen funktionellen Einheiten verglichen werden (z.B. varrierende Linker-Größe). Chemisch wenig sinnvolle MCSs können entstehen, wenn Ringe nur partiell zur Paarung gebracht werden können.[234] Hierdurch kann das Erkennen von Ähnlichkeiten verhindert und die SAR-Interpretation erschwert werden.

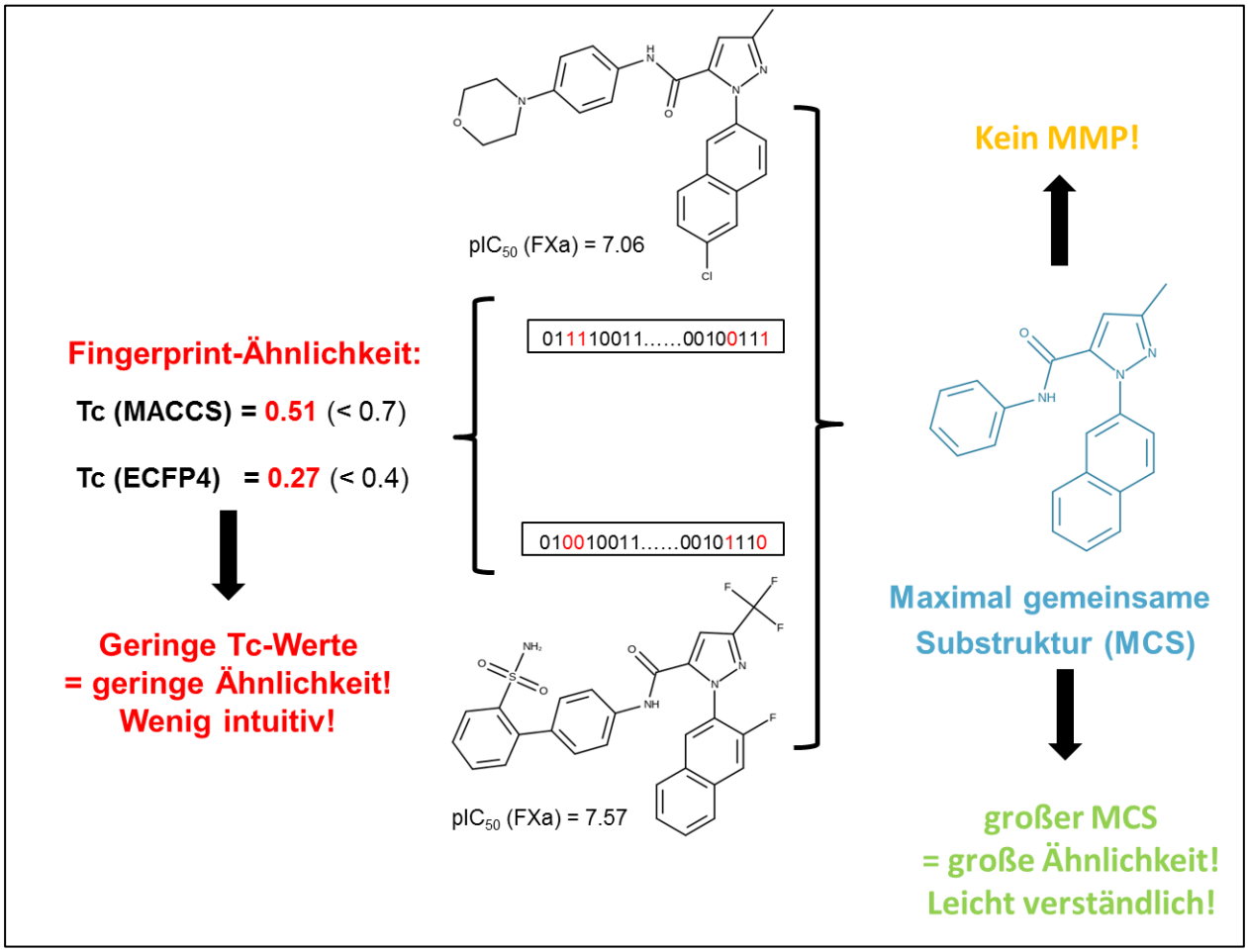


Abbildung 4.2. Vergleich des Konzeptes der Fingerprint-basierten Ähnlichkeit (am Beispiel der MACCS Keys und des ECFP4-Fingerprints), der maximalen gemeinsamen Substruktur (MCS, RASCAL Score = 0.50) und des zusammenpassenden Molekülpaares (MMP) am Beispiel von zwei Faktor Xa-Inhibitoren (FXa) mit vergleichbarer Bioaktivität. Für Details siehe Text.

4.4. Kombination mit Fingerprints

Da das Konzept des MCS zwar sehr intutiv, aber wenig geeignet für die Anwendung auf große Datensätze ist, findet man häufig Ansätze, bei denen Fingerprints und MCS kombiniert werden. Hierbei wird der Datensatz mit Hilfe von schnell berechneter Fingerprint-Ähnlichkeit geclustert und anschließend nur noch in den Clustern die rechenaufwändige Bestimmung des MCS durchgeführt. Hierdurch werden zeitaufwändige Vergleiche zwischen sehr unähnlichen Molekülen vermieden und eine Analyse großer Datensätze ermöglicht. Als Beispiele sind u.a. SHERIDAN[208], STAHL und MAUSER[229], BÖCKER[230], Chemaxon's LibMCS[231] und GARDINER et al.[232] zu nennen.

4.5. Vergleich mit dem MMP-Konzept

Die Konzepte des zusammenpassenden Molekülpaares (vgl. Abschnitt 2.6.5) und der maximal gemeinsamen Substruktur sind sehr ähnlich. Während beim MMP die gemeinsame Substruktur des Molekülpaares sehr groß ist und die strukturelle Änderung, die ein Molekül in das andere umwandelt, gering ist, kann der MCS zweier Moleküle von beliebiger Größe sein. Wenn der MCS eines Molekülpaares eine bestimmte Größe überschreitet, dann können die beiden Moleküle als MMP angesehen werden. Wie in Abschnitt 2.6.5 beschrieben, stellt die Bestimmung des MCS eine Möglichkeit dar, MMPs in einem Datensatz zu bestimmen. Aufgrund der Komplexizität der MCS-Berechnungen (v.a. mit steigender Datensatzgröße) werden oftmals aber andere Verfahren bevorzugt.

Der Nachteil von MMP-basierter im Vergleich zu MCS-basierten SAR-Analysen ist die Notwendigkeit der Definition einer maximalen Größe für das ausgetauschte Fragment (z.B. 10 Nicht-Wasserstoff-Atome bei REA und HUSSAIN[211]) und der Beschränkung der Zahl der ausgetauschten Fragmente. Dies schränkt die Flexibilität der Analyse unter Umständen deutlich ein. Das Molekülpaar aus Abbildung 4.2 würde demzufolge nicht als MMP identifiziert werden. Bei MCS-basierten Analysen ist es entscheidend, eine geeignete Mindest-MCS-Größe zu wählen bzw. die MCS-Größe immer in Relation zur ursprünglichen Größe der Moleküle zu betrachten. Die Erfordernis einer exakten Paarung der zu vergleichenden Moleküle stellt einen Nachteil beider Verfahren dar. Kleine strukturelle Unterschiede zwischen zwei Molekülen können verhindern, dass diese beiden Verfahren in der Lage sind, die vorhandene Ähnlichkeit zwischen den Molekülen zu erkennen.

5. Der Reduzierte Graph (RG)

5.1. Definition

Der Reduzierte Graph (RG) stellt eine Abstraktion des molekularen Graphen dar, indem er einzelne Atomgruppen zu funktionellen Einheiten zusammenfasst, die durch entsprechende Pseudoatome repräsentiert werden. Die ursprüngliche Molekül-Topologie wird hierbei erhalten. Je nach Abstraktionsgrad kann ein einzelnes Molekül durch verschiedene RG-Typen codiert werden. In den verschiedenen, publizierten RGs repräsentieren die einzelnen funktionellen Einheiten pharmakophore Merkmale oder physikochemische Eigenschaften. In diesem Fall können die RGs als eine Art „topologisches Pharmakophor“ betrachtet werden.[235] Im Gegensatz zu den klassischen 3D-Pharmakophoren wird hier jedoch das Problem der konformellen Flexibilität umgangen.

Einen guten Überblick über die Entwicklung der verschiedenen reduzierten Graphen, mit einem Schwerpunkt auf die Arbeiten an der Universität von Sheffield, und deren vielfältige Anwendungsmöglichkeiten geben BIRCHALL und GILLET[215] in einem Review.

5.2. Anwendung

Reduzierte Graphen wurden bisher für eine Vielzahl unterschiedlicher Aufgaben im Bereich der Chemoinformatik erfolgreich angewendet.[215] Eine der ersten Anwendungen war die Markush-Struktur-Repräsentation und die Suche in chemischen Patenten.[236]

Sehr erfolgreich wurden verschiedene Typen von RGs auch im Ligand-basierten VS mittels Ähnlichkeitssuche von verschiedenen Gruppen genutzt.[212, 235, 237–239] Hierbei hat sich gezeigt, dass sie eine gute Ergänzung zu den klassischen Fingerprints darstellen, die aufgrund ihres Prinzips eher strukturell sehr analoge Moleküle anreichern. RGs hingegen weisen aufgrund des Abstraktionsgrades ein deutliches höheres Scaffold-Hopping Potential auf.

Des Weiteren wurden RGs zum Clustern und der Analyse von HTS-Daten[201], sowie für die Identifizierung von repräsentativen Cluster-Molekülen verwendet[232]. Auch im Bereich der SAR-Analyse wurden RGs von BARKER et al. und BIRCHALL et al. (wie in Kapitel 2.6.6 ausführlich beschrieben) erfolgreich angewendet.[212–214] Ebenso konnten mit RGs erfolgreich bioisostere Austausche identifiziert werden.[216]

5.3. Verschiedene RG-Typen

RGs können auf verschiedene Weise dargestellt werden.[215] Sie können sowohl in Form von Fingerprints[212, 235] gespeichert werden, als auch als Graphen[232, 240] oder Pseudo-SMILES repräsentiert werden. Selbst RG-basierte SMARTS Ausdrücke (vgl. 2.6.6 und 10.2.3) sind möglich.[213–214] Die Graph-Repräsentation ermöglicht die Anwendung von

graphentheoretischen Algorithmen wie z.B. die Bestimmung der maximal gemeinsamen Substruktur und Substruktur-Suche. In RGs werden üblicherweise ungewöhnliche Atomtypen, die in normalen organischen Molekülen nicht vorkommen, als Pseudoatome zur Repräsentation verschiedener (pharmakophorer) Eigenschaften verwendet. Übergangsmetalle wie z.B. Sc, Cu, Co, Ni, Zn werden für diesen Zweck benutzt. Der Vorteil hiervon ist, dass Standard-Chemoinformatik-Software oder Programmierbibliotheken für die Erzeugung, Verarbeitung (z.B. für die MCS Bestimmung) oder Visualisierung der RGs verwendet werden können.

Das Entscheidende bei Verwendung von RGs ist die Wahl des richtigen Abstraktionsgrads, der immer in Abhängigkeit von der Problemstellung und dem Datensatz oder der Zielstruktur von Interesse festgelegt werden sollte. Wird das Molekül zu stark abstrahiert, werden die RGs zu unspezifisch. Es werden zwar unter Umständen große Ähnlichkeiten zwischen Molekülen gefunden, diese sind aber möglicherweise weder sinnvoll noch besonders informativ. Im umgekehrten Fall, wenn der gewählte Abstraktionsgrad zu niedrig ist bzw. durch die Pseudoatome sehr spezifische Eigenschaften codiert werden, ist es schwierig eigentlich vorhandene Ähnlichkeiten zu erkennen. Hier spiegelt sich erneut das in Kapitel 2.3 aufgezeigte Problem der adäquaten molekularen Repräsentation wider, dessen Lösung die Grundvoraussetzung für die angemessene Erfassung von chemischer Ähnlichkeit darstellt.

In der Literatur sind verschiedene Typen an reduzierten Graphen beschrieben. Hierbei kann man grob unterscheiden zwischen denen an der Universität von Sheffield entwickelten (GILLET et al.[235], BARKER et al.[212], BIRCHALL et al.[216]) oder damit eng verknüpften (HARPER et al.[201]) RGs, sowie den von RAREY und DIXON[237] entwickelten „Feature Trees“ und den „Erweiterten reduzierten Graphen“ (Abk. ErG), die auf STIEFL et al.[239] zurückgehen. Im Folgenden sollen die Hauptmerkmale nur kurz herausgestellt werden.

5.3.1. Klassische Sheffield-RGs

Die verschiedenen RGs unterscheiden sich vor allem in dem Abstraktionslevel und der Anzahl an codierten Eigenschaften bzw. Eigenschaftskombinationen, die durch ein RG-Pseudoatom repräsentiert werden. Die Codierung von Ringsystemen (aliphatisch, aromatisch, Annelierung) sowie die von terminalen Gruppen ist ebenfalls ein häufiges Unterscheidungskriterium.

Die meisten RG-Typen gehen auf die Arbeit von GILLET et al.[235] zurück bzw. stellen Erweiterungnen dieser dar. Bei diesem RG-Typ wurde bei der Molekülabstraktion der Schwerpunkt auf die Codierung von Ringsystemen und H-Brücken-Akzeptor- und -Donor-Eigenschaften gelegt, um funktionelle Gruppen hervorzuheben, die an Protein-Ligand-Interaktionen beteiligt sein können, und deren relative Position im Molekül darzustellen. Es wird zwischen einer R/F- und einer Ar/F-RG-Definition unterschieden. Bei der R/F-Abstraktion werden 3 Knotentypen definiert: Ring-Knoten (R), die sowohl aliphatische als auch aromatische Ringe repräsentieren, Merkmals-Knoten (F) mit H-Brücken-Bildungs-Eigenschaften und Linker-Knoten (L), die R- und F-Knoten miteinander verknüpfen. Bei der Ar/F-Abstraktion hingegen werden aliphatische Ringe nicht mehr explizit codiert und der allgemeine Ring-Knoten wird durch einen aromatischen Knoten (Ar) ersetzt. Aliphatische Ringe werden regelbasiert geöffnet und die entsprechenden Atome werden entweder als

Linker betrachtet, sofern Atome mit H-Brücken-Eigenschaften vorhanden sind als F-Knoten codiert oder aber bleiben wie alle terminalen, azyklischen Atome ohne HBA-/HBD-Eigenschaften unberücksichtigt.

BARKER et al.[212] haben zusätzlich zur R/F- und Ar/F-RG-Definition weitere Reduktionsschemata entwickelt. So werden im R/F/T-Schema terminale, azyklische Knoten (z.B. auch Methylgruppen) explizit als Knoten codiert, stellvertretend für hydrophobe Eigenschaften, die an wichtigen Protein-Ligand-Interaktionen beteiligt sein könnten. Auch wurde das R/F-Schema dahingehend modifiziert, dass bei den Ringen zusätzlich noch die Ringgröße als Merkmal berücksichtigt wird. Bei dem anschließenden Vergleich der verschiedenen RG-Definitionen im retrospektiven Virtuellen Screening war jedoch festzustellen, dass kein Abstraktionsniveau bzw. keine RG-Repräsentation gefunden werden kann, die für alle Target-Klassen zu optimalen Anreicherungen führt.

HARPER et al.[201] haben ausgehend von der Ar/F(4)-Definition von Gillet et al.[235] ein RG-Reduktionsschema entwickelt, bei dem sowohl weitere pharmakophore Eigenschaften (negativ und positiv ionisierbare Gruppen) hinzugefügt worden, als auch neben den aromatischen auch die aliphatischen Ringe als eigene Knotentypen codiert werden. Terminale, azyklische Gruppen ohne weitere pharmakophore Eigenschaften werden in dem Reduktionsprozess nicht berücksichtigt. Zudem wurde eine Prioritätsreihenfolge zur Entscheidung der Eigenschafts-Zuordnung (z.B. Positiv ionisierbar > H-Brücken-Donor) bei Erfüllung mehrer Eigenschaften festgelegt.

BIRCHALL et al.[216] unterscheidet sich von den vorgenannten Ansätzen dadurch, dass die Moleküle hierbei zunächst regelbasiert mittels SMARTS fragmentiert werden und den erhaltenen Fragmenten dann mittels SMARTS-Abgleich Eigenschaften zugeordnet werden. Diese annotierten Fragmente stellen im resultierenden RG die Knoten dar, die im Anschluss unter Erhalt der Topologie des ursprünglichen Moleküls wieder verbunden werden. Bei der Fragmentierung werden rekursiv alle nicht-terminalen, azyklischen Einfachbindungen geschnitten, wobei drei Ausnahmeregeln sicherstellen sollen, dass chemisch sinnvolle Fragmente erhalten werden. Bindungen zwischen zwei azyklischen sp^3-hybridisierten Kohlenstoffatomen (z.B. Alkylketten), Bindungen zwischen zwei azyklischen Heteroatomen und Bindungen zwischen einem azyklischen Heteroatom und einem azyklischen sp^2-hybridisierten Kohlenstoffatom (z.B. Carbonsäureester oder –amide) werden nicht gespalten. Bei der Zuordnung von Eigenschaften zu den einzelnen Fragmenten werden die folgenden Prioritätsreihenfolgen angewendet: aromatisch > aliphatisch > azyklisch und negativ ionisierbar > positiv ioinisierbar > HBAD > HBD oder HBA > keine Eigenschaft (keine H-Brückenbindungseigenschaft und keine Ladung).

5.3.2. ErG-Ansatz

Der von STIEFL et al. entwickelte ErG-Ansatz[239] weist große Ähnlichkeiten zu den in 5.3.1 dargestellten RG-Typen auf. Er stellt einen Hybrid-Ansatz aus RGs und Bindungs-Eigenschafts-Paaren (KEARSLEY et al.[241]) dar. Der Unterschied liegt insbesondere darin, dass STIEFL et al. versuchen die Größe eines Moleküls und seine räumliche Gestalt besser zu codieren und bestimmte Eigenschaften allgemeiner zu beschreiben. So werden

beispielsweise (anellierte) Ringe und Substituenten sowie terminale Gruppen („endcap groups“) im Vergleich zu BARKER et al.[212] alternativ codiert.

Bei ErG werden einzelnen Atomen gemäß des Protonierungszustand unter physiologischen Bedingungen Ladungen zugeordnet und funktionelle Gruppen mit HBD- und HBA-Eigenschaften annotiert. Es werden nur terminale hydrophobe Gruppen mit drei Atomen (z.B. Isopropylgruppen) und Thioether behalten. Diese sogenannten hydrophoben „endcap groups“ sollen dazu dienen, die Größe und räumliche Gestalt des Moleküls besser abzubilden. Eine weitere Besonderheit stellt die Codierung von Ringsystemen dar. Ringe mit einer Ringgröße kleiner oder gleich sieben Ringatomen werden codiert, Makrozyklen bleiben unverändert. Für jeden Ring wird ein Centroidatom hinzugefügt, das, sofern der Ring aus mehr als 50% sp^2-hybridisierten Atomen besteht, mit der Eigenschaft „aromatisch“ annotiert wird. Auch wenn es sich um einen direkt an einen aromatischen Ring annelierten aliphatischen Ring handelt, bekommt sein Centroid nicht die Eigenschaft „hydrophob“, sondern „aromatisch“ zu gewiesen, um der reduzierten Flexibilität und verstärkten Hydrophobizität Rechnung zu tragen. Die Centroidatome werden mit allen substituierten Ringatomen und allen Ringatomen, denen eine pharmakophore Eigenschaft zugewiesen wurde, sowie allen Brückenkopfatomen verbunden. Alle Ringatome, die nicht mit einem Centroidatom verbundenen sind, werden entfernt. Im Gegensatz dazu werden zur Codierung von Entfernungen zwischen den einzelnen pharmakophoren Eigenschaften alle Nicht-Ringatome, auch wenn sie keine Interaktionseigenschaften (Linker-Atome) aufweisen, erhalten.

5.3.3. Feature Trees

Bei dem Feature Tree (Abk. FT) Ansatz von RAREY und DIXON[237] wird ein Molekül als Eigenschafts-Baum repräsentiert. Jeder FT-Knoten stellt ein Molekülfragment dar und Kanten verbinden zwei Knoten, sofern die repräsentierten Fragmente im ursprünglichen 2D-Molekülgraphen verbunden waren. Das Besondere bei den FTs im Gegensatz zu anderen RG-Typen ist, dass regelbasiert sichergestellt wird, dass FTs immer Bäume sind (also niemals Zyklen enthalten), um den algorithmischen Aufwand beim Vergleich von FTs gering zu halten.

Bei der Erzeugung von FT-RGs werden dafür zunächst alle Ringsysteme identifiziert und solange zu Knoten zusammengefasst bis keine zyklischen Strukturen mehr auftreten. Bei den nicht-zyklischen Molekülteilen stellt zunächst jedes Atom, das mehr als eine Bindung aufweist, einen Einzelknoten im FT dar. Die terminalen Atome hingegen werden mit den verbundenen Atomen zu einem Knoten zusammengefasst, wobei Wasserstoffatome auch als Atome berücksichtigt werden, sodass z.B. Hydroxyl- oder Aminogruppen in jeweils einem Knoten zusammengefasst werden. Für die Bestimmung der Ähnlichkeit zwischen zwei Bäumen werden zunächst paarige Sub-Bäume identifiziert und dann hierfür eine gewichtete Merkmals-Ähnlichkeit für die zusammenpassenden Knoten berechnet. Durch Zusammenfassen einzelner Sub-Bäume zu einem Knoten lässt sich der Grad an Spezifität der FTs variieren.

Im Gegensatz zu anderen RGs gibt es bei dem FT-Ansatz zwei Eigenschaftsklassen, d.h. dass die Knoten nicht nur mit pharmakophoren, sondern auch sterischen Eigenschaften

annotiert werden. Zur Beschreibung der Größe der einzelnen Fragmente wird zum einen die Anzahl an Fragment-Atomen, zum anderen das approximierte van-der-Waals Volumen berücksichtigt. Die räumliche Gestalt wird damit jedoch nicht beschrieben, da diese in der Regel abhängig von der molekularen Konformation ist, die bei diesem Ansatz nicht berücksichtigt werden soll. Für die Zuordnung von chemischen Eigenschaften wird für jedes Fragment ein Interaktionsprofil bestimmt. Dafür kann sowohl das mittels Interaktionsenergie gewichtete FlexX Interaktionsprofil[242] (Unterscheidung der folgenden Interaktionstypen: H-Brücken-Donor, H-Brücken-Akzeptor, Metall-Interaktion, aromatisches Ringzentrum, aromatisches Ringatom, Methylgruppe, Amid und hydrophob) als auch ein nicht-gewichtetes Profil basierend auf verschiedenen Atomtypen verwendet werden.

5.4. Abgrenzung zu Topologischen Pharmakophor-Deskriptoren

Ligand-basierte topologische Pharmakophore sind den RGs sehr ähnlich. Hierbei handelt es sich sich um eine Gruppe von Deskriptoren, bei denen man versucht die 3D-Pharmakophor-Repräsentation auf 2-dimensionaler Ebene zu imitieren. Hierfür werden einzelne Atome zu pharmakophoren Eigenschaften bzw. funktionellen Atomtypen abstrahiert und durch die Erfassung von Distanzen zwischen diesen Atomen im 2-dimensionalen chemischen Graphen wird versucht, die geometrischen Distanzen im Raum nachzubilden.[243]

Der Unterschied zu RGs besteht darin, dass bei RGs ganze funktionelle Gruppen zu pharmakophoren Pseudoatomen abstrahiert werden, während bei den topologischen Pharmakophor-Deskriptoren nur eine Abstraktion einzelner Atome zu funktionellen Atomtypen stattfindet.[243] Durch diese stärkere Abstraktion der RGs sind sie weniger abhängig von der Topologie einzelner funktioneller Gruppen und können so unter Umständen Ähnlichkeiten erkennen (z.B. aufgrund unterschiedlich langer Linker, der im RG nur durch ein Linker-Pseudoatom codiert wird), die den topologischen Pharmakophoren verborgen bleiben. Jedoch besteht bei den RGs die Gefahr der zu hohen Abstraktion.

Ein Beispiel für einen topologischen Pharmakophor-Dekriptor ist der von SCHNEIDER et al. entwickelte CATS2D (Chemically Advanced Template Search) Deskriptor[175, 244], für den eine erhöhte Anreicherung von diversen Scaffolds („Scaffold-Hopping“) im VS gezeigt werden konnte[245]. Bei CATS2D wird ein Atom eines Moleküls SMARTS-basiert zu einem pharmakophoren Atomtyp (H-Brücken-Akzeptor, H-Brücken-Donor, positiv gelanden, negativ geladen und lipophil) abstrahiert und anschließend wird zwischen all diesen Atompaaren der kürzeste Pfad bestimmt. Die Atompaare werden bis zu einer bestimmten Distanz in einem Korrelationsvektor gespeichert.

Weitere Beispiele sind der Similog Deskriptor[246] von SCHUFFENHAUER et al., der speziell für die Identifizierung von bioisosterem Austausch entwickelte Atompaar-Deskriptor von WAGENER und LOMMERSE[247] oder der von HOLLIDAY et al. entwickelte R-Gruppen-Deskriptor[248] zum Vergleich von (nicht-)bioisosteren Substituenten basierend auf der Verteilung von Atom-basierten physiokochemischen Eigenschaften.

5.5. Synergismus des MCS-Konzeptes und RGs

Wenn man RGs wie mathematische Graphen behandelt, hat es den Vorteil, dass graphentheoretische Algorithmen wie die Bestimmung des MCS oder Substruktur-Suchen verwendet werden können (vgl. GARDINER et al.[232] und BARKER et al.[240]).

Wie mehrfach dargestellt, ist das Konzept des MCS sehr anschaulich, jedoch auch mit einigen Nachteilen verbunden (vgl. Abschnitt 4.3). Aufgrund der Notwendigkeit der exakten Atompaarung resultieren manchmal beispielsweise chemisch wenig sinnvolle MCSs (z.B. unvollständige Ringsysteme). Der MCS (rot markiert) in Abbildung 5.1, der aus dem Vergleich der beiden fast äquipotenten COX2-Inhibitoren (links) resultiert ist ein Beispiel für diesen Fall (vgl. jeweils den mittleren Ring in den Molekülen). Anhand des MCS der Moleküle kann das für viele COX2-Inhibitoren typische Motiv aus 3 Aromaten und H-Brücken-Akzeptor (sogenanntes „Mickey-Maus“ Motiv, vgl. Abschnitt 20.2.2) nicht abgeleitet werden. Abstrahiert man jedoch die Moleküle zu RGs (wie rechts in Abbildung 5.1 dargestellt), kann man sehen, dass das Motiv (lila markiert) in dem aus dem Vergleich beider RGs resultierenden RG-MCS (rot) codiert ist. Zudem bestehen die RGs im Vergleich zu den Molekülen nur aus wenigen Pseudoatomen, sodass die Komplexität der Berechnungen deutlich reduziert ist.[232] Durch die Kombination des MCS-Konzeptes mit RGs resultieren somit einige Vorteile bzw. ist es möglich einige der oben erwähnten Nachteile zu beheben.

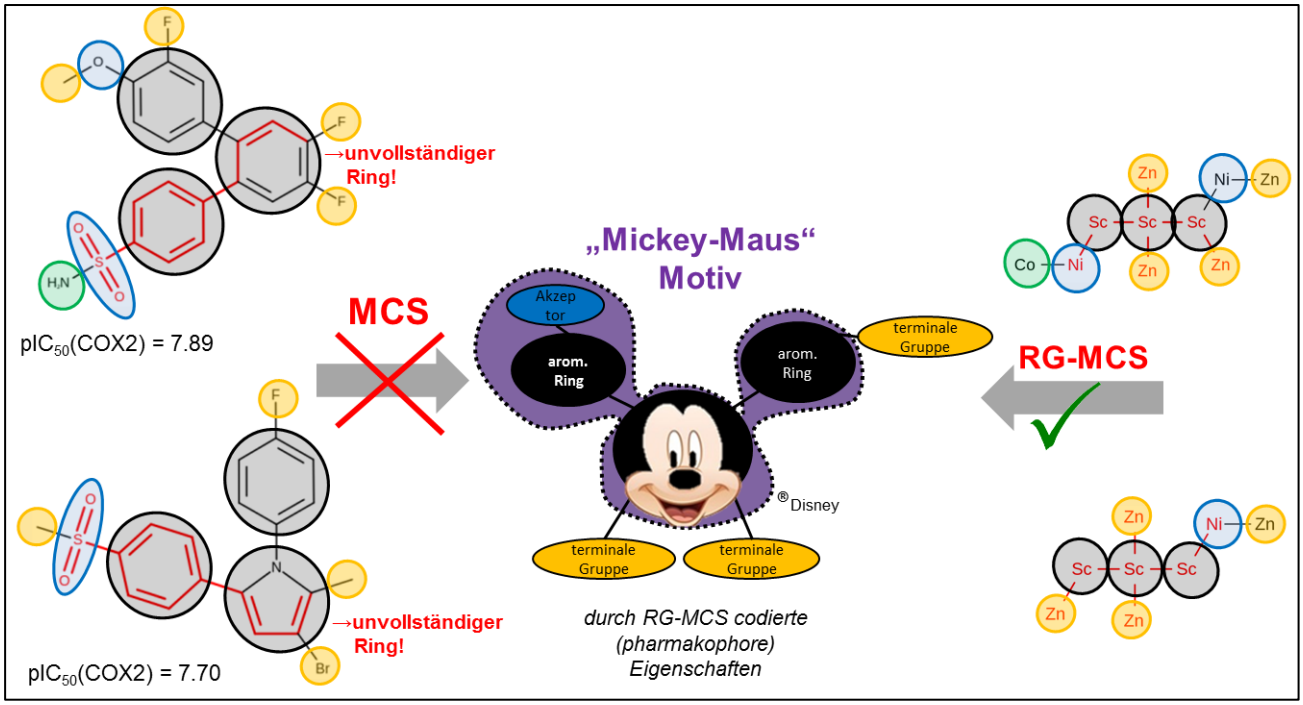


Abbildung 5.1. Veranschaulichung des Vorteils der aus der Kombination des MCS-Konzeptes mit RGs resultiert am Beispiel des für Cyclooxygenase 2 (COX2) Inhibitoren typischen pharmakophoren „Mickey-Maus“-Motivs (lila markiert). Details siehe Text.

6. Nicht-kovalente Protein-Ligand-Interaktionen

Protein-Ligand-Interaktionen stellen wie unter 2.2 beschrieben eine wichtige Grundlage für biologische Aktivität an Protein-Targets dar. Die Interaktion eines Moleküls mit seiner Zielstruktur kann dabei sowohl auf der Ausbildung einer kovalenten Bindung, als auch auf nicht-kovalente Wechselwirkungen zurückzuführen sein. Kovalentes Binden (z.B. zytostatischer Alkylantien) hat oftmals den Nachteil mangelnder Selektivität und Reversibilität, einhergehend mit vielfältigen toxischen unerwünschten Arzneimittelwirkungen. Spezifische kovalente Inhibitoren (z.B. Protonenpumpen-Inhibitoren) können jedoch den Vorteil u.a. einer längeren Wirkdauer haben. Die Wirkung der Mehrheit an Arzneistoffe beruht auf nicht-kovalenten Protein-Ligand-Wechselwirkungen. Da diese die Grundlage für die SAR-Analysen mittels der in dieser Arbeit entwickelten inSARa-Methode darstellen, gibt der folgende Abschnitt einen kurzen Überblick über die wichtigsten nicht-kovalenten Protein-Ligand-Wechselwirkungen. Für weiterführende Literatur sei auf eine Vielzahl guter Übersichtsartikel[42, 249–251] und Bücher[34, 44, 252] verwiesen.

6.1. Ionische Wechselwirkungen

Ionische Interaktionen gehören wie die H-Brücken-Bindungen zu den elektrostatischen Wechselwirkungen, sind jedoch deutlich seltener in Protein-Ligand-Komplexen zu finden.[34] Sie beruhen auf dem Prinzip, dass eine geladene Gruppe des Liganden mit einer entgegengesetzt geladene Aminosäure-Seitenkette des Proteins in Wechselwirkung tritt.[34] Im Gegensatz zu H-Brücken-Bindungen sind sie ungerichtet, über einen größeren Radius wirksam und in der Regel deutlich stärker (Beiträge zur Bindungsenergie im Vakuum zu Grunde gelegt, in wässrigem Medium aufgrund des Desolvatationseffektes meist mit neutralen H-Brücken vergleichbar)[250]. Da ionische Interaktionen häufig gemeinsam mit H-Brücken-Bindungen auftreten, ist eine klare Abgrenzung beider Interaktionen daher häufig schwierig.[250] Ionische Interaktionen weisen aufgrund des Einflusses auf den Ladungszustand des Liganden eine starke Abhängigkeit von dem umgebenden Medium (Dielektrizitätskonstante, pH-Wert) bzw. dem pK_a-Wert der basischen oder sauren funktionellen Gruppe auf.[250] Schwierigkeiten bei der Bestimmung des Protonierungszustandes des Liganden sowie schwer vorhersagbare Solvatations- und Desolvatationseffekte sind dafür verantwortlich, dass sich die Bedeutung von ionischen Wechselwirkungen für die Bindungsaffinität im Protein-Ligand-Komplex nur schwer abschätzen lässt. Zahlreiche Literaturbeispiele[42, 251] (z.B. Inhibitoren von Koagulations-Faktor Xa oder Thrombin) zeigen jedoch, dass ionischen Wechselwirkungen essentielle Schlüssel-Interaktionen an einer Zielstruktur darstellen.

6.2. Wechselwirkungen mittels H-Brücken-Bindungen

H-Brücken-Bindungen beruhen auf einer Dipol-Dipol-Interaktion zwischen einem H-Brücken-Akzeptor (Abk.: HBA) und -Donor (Abk.: HBD) und sind ebenso elektrostatischer Natur.[253-254] Sie sind jedoch gerichtet und über einen kürzeren Radius wirksam.[253–254] Eine

H-Brücken-Bindung kann sich zwischen einem positiv polarisierten Wasserstoff-Atom (HBD), das kovalent an ein elektronegativeres Atom gebunden ist, und einem weiteren elektronegativen Atom mit freiem Elektronenpaar (HBA) ausbilden, sofern es zu einem Gewinn an freier Energie kommt.[253–254] Diese Änderung in der freien Energie wird stark beeinflusst von Faktoren wie der räumlichen Geometrie, der Zugänglichkeit der Bindungsstelle für das wässrige Medium und den Eigenschaften der benachbarten Atome.[42] Ist einer der Bindungspartner zusätzlich noch geladen, handelt es sich um eine *ladungsunterstützte* (durch die Ladung verstärkte) *H-Brücken-Bindung*.[250] Im Vakuum weisen diese eine deutlich höhere Bindungsenergie als neutrale H-Brückenbindungen auf. Im wässrigen Medium liegt ihr Beitrag zur Bindungsenergie aufgrund der erschwerten Desolvatation i.d.R. im Bereich einer normalen Wasserstoffbrücke (-2 bis -6 kJ/mol).[250]

Wasserstoffbrücken spielen eine entscheidende Rolle für die Orientierung, die spezifische molekulare Erkennung (im Gegensatz zu hydrophoben Interaktionen) und die Affinität eines Liganden im Protein-Ligand-Komplex.[255] Ein bisher ungelöstes Problem ist jedoch, dass ihr Beitrag im Protein-Ligand-Komplex aufgrund der oben genannten vielfältigen Einflussfaktoren schwer vorhersagbar ist.[42, 256] Während die Bindungsgeometrie (Distanzen, Winkel) aufgrund von statistischer Auswertung (u.a. durch IsoStar[257], SuperStar[258] und Relibase[259]) der zunehmenden Anzahl an verfügbaren Kristall-Strukturen von Protein-Ligand-Komplexen in der PDB und CSD[260] gut beschrieben werden kann, ist die Bindungsenergie bzw. die Stärke der Bindung v.a. aufgrund des schwer berechenbaren Einflusses von Solvatations- und Desolvatationseffekten nur schwer abschätzbar.[42, 250] Denn bei der Ausbildung einer Wasserstoffbrücke steht der Ligand immer im Wettbewerb mit Wassermolekülen und die Desolvatation der solvatisierten Bindungsstelle und des solvatisierten Liganden stellt eine essentielle Voraussetzung für das Ausbilden der Bindung dar. Das bedeutet, dass, obwohl die Ausbildung geometrisch günstig wäre, es nicht unbedingt zu einer Interaktion kommt bzw. dass diese Interaktion nicht zwingend einen hohen Beitrag zur Bindungsaffinität des Liganden leistet. Obwohl grob eine Zunahme der Bindungsaffinität um eine Größenordnung pro H-Brückenbindung festzustellen ist, ist keine Korrelation der Anzahl an H-Brückenbindungen in einem Komplex mit der Bindungsaffinität möglich.[249] Es sind zahlreiche Protein-Ligand-Komplexe beschrieben, die trotz fehlender oder nur weniger H-Brücken-Bindungen nanomolare Affinität aufweisen.[249] H-Brücken sind v.a. für die Spezifität eines Komplexes wichtig.[256]

Neben den klassischen Wasserstoffbrücken haben in den letzten Jahren auch die bisher wenig beachteten, sogenannten „*schwachen H-Brücken-Bindungen*“ (unter Beteiligung von CH-Gruppen, Fluor-Atomen oder π-Elektronen) immer öfter Erwähnung in der Literatur gefunden.[261–262] Hier ist es jedoch noch schwieriger die Bedeutung für die Bindungsaffinität richtig abzuschätzen.[42]

6.3. Hydrophobe Interaktionen

Während der Beitrag von spezifischen H-Brücken-Interaktionen zur Bindungsaffinität schwer abzuschätzen ist und meist überschätzt wird, wurde der Einfluss unspezifischer, hydrophober Wechselwirkungen aufgrund der ungerichteten diffusen Natur lange Zeit unterschätzt.[256] Es ist jedoch festzustellen, dass hydrophobe Interaktionen einen Hauptbeitrag zur Ligand-Affinität in vielen Protein-Ligand-Komplexen leisten.[256] Der

Hauptanteil der Bindungsenergie entsteht nicht durch die schwachen (Van-der-Waals-) Anziehungskräfte der hydrophoben Gruppen, sondern durch den entropischen Desolvatationsbeitrag (sogenannten „hydrophoben Effekt“)[249, 256]. Dieser beruht auf dem Prinzip, dass die geordnete Wassernetzwerk-Struktur durch lipophile Liganden zerstört wird, wodurch ein Entropie-Gewinn resultiert.[42, 249] Außerdem können Wassermoleküle mit lipophilen Bindetaschen keine H-Brücken-Bindungen eingehen.[249] Durch die Verdrängung ist nun H-Brücken-Bindung mit anderen Wassermolekülen möglich, woraus zusätzlich noch ein enthalpischer Beitrag resultiert.[249]

Es existiert eine positive Korrelation zwischen der lipophilen Kontaktoberfläche und der Bindungsaffinität.[249] Die Stärke des hydrophoben Effektes kann mit einer etwa 3,5-fachen Zunahme der Bindungskonstante pro Methylgruppe abgeschätzt werden.[42]

6.4. Aromatische π-π–Interaktionen

Abgesehen von ungerichteten hydrophoben Interaktionen können aromatische Ringe auch spezifische, gerichtete π-π-Interaktionen (meist T-förmige „face-to-edge“ oder parallel verschobene „face-to-face“ Anordnung) eingehen, was auf die große Polarisierbarkeit und das beachtliche Quadrupolmoment des π-Elektronensystems zurückzuführen ist.[263–264] In Protein-Ligand-Komplexen sind π-π–Interaktionen häufig zu finden und können einen wichtigen Beitrag zur Bindungsaffinität eines Liganden leisten.[42]

6.5. Weitere Protein-Ligand-Wechselwirkungen

Kationen- π-Interaktionen

Kationen-π-Interaktionen stellen oftmals wenig beachtete, wichtige Wechselwirkungen an einer Vielzahl von Targets (z.B. Acetylcholinesterase, nikotinischer Acetylcholinrezeptor, muskarinischer Acetylcholinrezeptor und viele weitere GPCRs wie z.B. der D_2-Rezeptor) dar.[265–266] Die Interaktion kann als elektrostatische Anziehung zwischen der positiven Ladung und dem Quadrupolmoment des Aromaten betrachtet werden.[265] Aromatische Aminosäuren haben somit einen polaren als auch einen hydrophoben Charakter, sodass durch Aromaten gebildete, hydrophobe Bindestellen in der Lage sind auch polare, kationische Liganden zu binden (z.B. in der S4-Tasche bei Faktor Xa).[265, 267]

Wechselwirkungen mit Halogen-Atomen

Halogen-Atome sind häufiger Bestandteil von Arzneistoffen, beispielsweise enthalten etwa 20-25% aller auf dem Markt befindlichen Arzneistoffe mindestens ein Fluor-Atom, darunter auch viele Blockbuster (z.B. Atorvastatin, Ciprofloxacin, Celecoxib, Fluoxetin, Sitagliptin, Evavirenz).[268–269] Daher sollte man, wenn man SARs analysiert, hinterfragen, ob bzw. welche Rolle sie bei der Interaktion des Liganden mit dem Target spielen.

Eine spezifische Interaktion der Halogen-Atome, die viele Analogien zur H-Brücken-Bindung aufweist, ist die *Halogen-Bindung*.[42, 270–272] Eine Halogen-Bindung (C–X···Y) ist eine elektrostatische Interaktion zwischen einem Halogen-Atom (X = Cl, Br, I) und einer Lewis-Base (Y), in den meisten verfügbaren Protein-Ligand-Komplexen das freie Elektronpaar des Sauerstoffatoms einer Carbonylgruppe oder den π-Elektronen eines elektronenreichen Aromaten. Die Polarisation des Halogen-Atoms entlang der C–X σ-Bindung verursacht eine anisotrope Ladungsdichte-Verteilung an dem Halogen-Atom. Dies führt zu einem Elektronendefizit, auch σ-Loch genannt, auf der gegenüberliegenden Seite der σ-Bindung, das kompensatorisch von einem elektronenreichen Ring umgeben ist und das der Grund dafür ist, dass das Halogenatom als Elektronen-Akzeptor interagiert und die resultierende nicht-kovalente Bindung stark entlang der Achse der σ-Bindung gerichtet ist. Die Stärke der Bindung nimmt mit zunehmender Größe des Halogenatoms (I > Br > Cl) aufgrund der größeren Polarisierbarkeit der Elektronen zu. Halogen-Bindungen sind dennoch deutlich schwächer als normale H-Brückenbindungen. Die Existenz von Halogen-Bindungen ist schon lange bekannt (Nobelpreis 1969)[273], ihre genaue Ursache, die Häufigkeit ihres Vorkommens sowie ihre Bedeutung für die Stabilität von Protein-Ligand-Komplexen wird hingegen erst in den letzten Jahren immer offensichtlicher[271].

Fluor nimmt aufgrund seiner hohen Elektronegativität und geringeren Polarisierbarkeit eine Sonderstellung unter allen Halogen-Atomen ein und geht i.d.R. *keine* Halogen-Bindung ein.[42] Stattdessen zeigen Fluor-Atome einen amphiphilen Charakter und können je nach lokaler chemischer Umgebung entweder als elektronenreicher Akzeptor mit einem H-Brücken-Donor interagieren oder aber hydrophobe Interaktionen mit lipophilen Aminosäure-Seitenketten und sogenannte *orthogonale multipolare Interaktionen* mit dem Kohlenstoff-Atom von Carbonylgruppen oder dem Guanidin des Arginins eingehen.[274–275] Zudem sind Fluor-Substituenten stark elektronenziehend, wodurch sie die Acidität (pK_a) benachbarter funktioneller Gruppen stark beeinflussen, was wiederum vielfältige Auswirkungen auf die Eigenschaften eines Arzneistoffes hat (wie HAGMANN[269] und MÜLLER et al.[276] sehr ausführlich aufzeigen).

Komplexbildung mit Metallionen

Die Wechselwirkung zwischen Metallionen und Liganden stellt einen weiteren speziellen Interaktionstyp dar[277–278], der v.a. in Metalloenzymen von Bedeutung ist, d.h. katalytisch aktive Metalloproteine, in denen ein bestimmtes Metallion im aktiven Zentrum an der katalytischen Reaktion beteiligt ist. Eines der häufigsten Metallionen in Metalloenzymen stellt das Zink dar. Es findet sich beispielsweise im katalytischen Zentrum bei den Angiotensin-Konversions-Enzymen, den Matrix-Metalloproteasen, sowie den Carboanhydrasen oder aber in Kombination mit Magnesium bei den Phosphodiesterasen.[34]

Sogenannte „Zink-bindende Gruppen“ (Abk. ZBG) in Liganden können Komplexe mit Metall-Ionen in aktiven Zentren bilden und leisten so u.U. einen Beitrag zur Affinität des Liganden. ZBG sind strukturell sehr divers (z.B. Thiolgruppen, Hydroxamsäuren, Carboxylate, Sulfonamide[34, 279]) und ihre Affinität ist stark vom Target abhängig[34]. Die Vorhersage dieser Interaktion bzw. Modellierung ist daher sehr schwierig.[280–282]

7. Pharmakophore Eigenschaften

Basierend auf den Erkenntnissen zu den am häufigsten vorkommenden nicht-kovalenten Protein-Ligand-Interaktionen (vgl. Kapitel 6) werden typischerweise die folgenden pharmakophore Eigenschaften unterschieden: Ionische Eigenschaften (negativ oder positiv ionisierbares Zentrum), H-Brücken-Akzeptor- und Donor-, sowie hydrophobe und aromatische Eigenschaften.[283]

Die Verwendung dieser Eigenschaft hat sich in einer Vielzahl von Studien als erfolgreich für das Erkennen und das Verständnis von Schlüssel-Interaktionen zwischen einem Target und seinen Liganden herausgestellt.[283] Einen Überblick über das Pharmakophor-Konzept und eine Vielzahl von Anwendungsbeispielen in der medizinischen Chemie (Virtuelles Screening und Erklären von SARs) mit dem Schwerpunkt auf 3D-Pharmakophore, sowie Verweise auf weiterführende Literatur geben zwei Bücher[284–285]. Des Weiteren sei auf einen guten Übersichtsartikel von LEACH et al.[286] verwiesen.

Eine grundlegende Arbeit zur Definition von pharmakophoren Eigenschaften geht auf GREENE et al.[287] zurück (implementiert im kommerziellen Programm Catalyst[288]). Sie bildet die Grundlage für viele weitere Arbeiten (z.B. das TAMINAU et al.[289]) bzw. viele Programme zum Ableiten von Pharmakophoren beruhen auf ähnlichen Prinzipien (vgl. Übersicht von WOLBER et al.[290]).

Im Folgenden werden die wichtigsten Aspekte, die bei der Definition von pharmakophoren Eigenschaften zu berücksichtigen sind, zusammengefasst. Da die Analysen in dieser Arbeit basierend auf 2-dimensionalen Molekülstrukturen durchgeführt werden, werden Aspekte, die nur 3D-Pharmakophore betreffen (z.B. räumliche Platzierung von pharmakophoren Sphären, Sphärendurchmesser, Direktionalität von Eigenschaften in Form von projizierten Punkten[290]) nicht weiter betrachtet.

7.1. Ionische Eigenschaften

Nach GREENE et al.[287] lassen sich ionische Eigenschaften wie folgt definieren: Atome, die unter Berücksichtigung des Protonierungszustand bei physiologischen pH-Wert (7,4) eine negative oder positive Formal-Ladung tragen werden als positiv (Abk. PI) oder negativ ionisierbare (Abk. NI) Zentren betrachtet, sofern sie nicht direkt zu einem entgegengesetzt geladenen Atom benachbart sind (z.B. Nitrogruppe). Zusätzlich sind delokalisierte Ladungen aufgrund von delokalisierten π-Elektronen-Systemen zu berücksichtigen (z.B. Carboxyl-Gruppe, Tetrazol oder Guanidin). In vielen Fällen wird die Eigenschaft im Centroid der Heteroatome der funktionellen Gruppe platziert.

Typische funktionelle Gruppen, die NI darstellen, sind[287, 291]: Carbonsäure, S-/P-Säuren, Tetrazol, Imid, Sulfonylcarbamid, Trifluormethylsulfonamid. Typische funktionelle Gruppen, die PI darstellen, sind[217, 221]: Aliphatische Amine, Amidin, Guanidin. Darüberhinaus gibt es noch schwächere Basen und Säuren, deren Protonierungszustand, stark von weiteren funktionellen Gruppen in der Nachbarschaft abhängig ist[287] (z.B. 2-/4-Aminopyridine[291], Imidazole).

Obwohl viele Pharmakophor-basierte Ansätze PI und NI explizit codieren, ist es umstritten, ob dies aufgrund der unvorhersagbaren Solvatations- und Desolvationseffekte sinnvoll ist oder stattdessen nur mit H-Brücken-Akzeptor- und Donor-Eigenschaften gearbeitet werden sollte.[286] Der Protonierungszustand ist dabei aber dennoch von entscheidender Bedeutung (z.B. unprotoniertes aliphatisches Amin = H-Brücken-Akzeptor, protoniertes Amin = H-Brücken-Donor).[292] Zudem ist zu beachten, dass unterschiedliche tautomere Formen unterschiedliche pK_a-Werte besitzen können bzw. Tautomerie auch entscheiden Einfluss auf die Erkennung von H-Brücken-Akzeptoren/-Donoren und Aromatizität hat.[293–294]

7.2. H-Brücken-Akzeptor Eigenschaften

H-Brücken-Akzeptor-Eigenschaften (Abk. HBA) lassen sich nach GREENE et al.[287] wie folgt definieren: Jedes N-, O-, S-Atom mit mindestens einem verfügbaren (d.h. nicht-delokalisiertem) freiem Elektronenpaar wird als HBA betrachtet. Basische Amine werden aufgrund von Protonierung als HBA ausgeschlossen.
Oftmals wird der Ausschluss einiger Atome in bestimmter chemischer Umgebung als HBA diskutiert.[286] Hierdrunter fallen beispielsweise einige O-Atome (z.B. solche in Furan- oder Oxazol-Ringen), bei denen es sich bekanntermaßen um schwache Akzeptoren handelt.[42, 295] Ebenso ist zu beachten, dass der sp^3-hybridisierte Ester-Sauerstoff im Gegensatz zum Carbonyl-Sauerstoff, wie Analysen der CSD zeigen, zumeist nur ein sehr schwacher HBA ist. KUBINYI verweist darauf, dass falsche Erkennung und Zuordnung von Eigenschaften oftmals die Ursache schlechter Ergebnisse von Pharmakophor-Suchen oder Docking ist.[292]

TAMINAU et al.[289] ergänzen die obige HBA-Definition, um die folgenden Ausnahmen: Alle N-Atome oder O-Atome, sofern sie keine positive Formalladung aufweisen, mindestens ein freies, nicht-delokalisiertes Elektronenpaar besitzen und sterisch zugänglich sind, werden als HBA erkannt. Dabei werden nur die N-Atome berücksichtigt, die weniger als drei Verbindungen besitzen, sofern sie Teil eines aromatischen Ringsystems sind (Ausschluss z.B. von Pyrrol-Stickstoff, aber nicht von Pyridin-Stickstoff, da das freie Elektronenpaar beim Pyrrol-N Bestandteil des aromatischen Systems ist und somit nicht für die H-Brücken-Bildung zur Verfügung steht). Auch werden direkt an einen aromatischen Ring gebundene Stickstoffatome, wenn sie drei Verbindungen aufweisen (Anilin-Stickstoff), ausgeschlossen. Dasselbe gilt für (Sulfon-)Amid-Stickstoff-Atome, da auch hier eine Delokalisierung des freien Elektronenpaars sehr wahrscheinlich ist. Sterische Zugänglichkeit bedeutet, dass genug Platz zur Ausbildung der H-Brücken-Bindung vorhanden ist und keine sterische Hinderung durch andere Molekülteile zu erwarten ist. Folglich ist dieses Kriterium abhängig von der vorliegenden Konformation. Zur Prüfung werden im Radius von 1.8 Å um das potentielle H-Brücken-Akzeptor-Atoms Punkte gleichmäßig im Raum verteilt. Wenn mindestens 2% dieser Punkte nicht mit irgendwelchen Nachbaratomen kollidieren, wird das entsprechende Atom als H-Brücken-Akzeptor eingestuft.

7.3. H-Brücken-Donor Eigenschaften

H-Brücken-Donor-Eigenschaften (Abk. HBD) werden nach GREENE et al. wie folgt definiert[287]: Alle nicht sauren Hydroxylgruppen, alle Stickstoffatome mit gebundenem

Wasserstoffatom, sofern sie nicht Bestandteil von sauren funktionellen Gruppen (Tetrazol oder Trifluormethylsulfonamid) sind, sowie Thiol- oder Acetylengruppen werden als HBD betrachtet. Dabei sind die beiden letztgenannten aufgrund sehr schwacher Donor-Eigenschaften in Literatur umstritten.[42, 286] Je nach betrachteter Zielstruktur ist es jedoch sinnvoll diese schwachen und weitere schwache HBD (z.B: aromatische CH-Gruppen in manchen Kinase-Inhibitoren[296]) als HBD zu definieren. TAMINAU et al.[289] berücksichtigen nur O- und N-Atome als HBD, an die mindestens ein Wasserstoffatom gebunden ist und die keine negativen Formalladungen aufweisen.

7.4. Hydrophobe Eigenschaften

Die Definition von hydrophoben Eigenschaften ist nicht so einfach und einheitlich wie die der vorangehend beschriebenen Eigenschaften (vgl. LEACH et al.[286]). Pharmakophor-Programme unterscheiden sich deutlich in der Definition hydrophober Eigenschaften wie WOLBER et al. zeigen konnten.[290] Selbst bei Programmen, die auf demselben Algorithmus beruhen, sind deutliche Unterschiede festzustellen.[297]

Ein grundlegender Algorithmus zur Definition von Hydrophobizität, der die Grundlage für viele 3D-Pharmakophor-Programme darstellt[288–289, 298–299], geht ebenfalls auf GREENE et al. zurück[287]. Das Prinzip beruht darauf, dass jedem Atom regelbasiert ein von seiner direkten Nachbarschaft und dem Atomtyp abhängigiger tabellierter Hydrophobizitätswert zugewiesen wird, der auf empirische Beurteilung medizinischer Chemiker zurückgeht. Zusätzlich zu dem Hydrophobizitätswert wird noch der Anteil an Lösungsmittel zugänglicher Oberfläche für jedes Atom berücksichtigt. Im letzten Schritt werden benachbarte Atome zu kleinen Gruppen geclustert, wobei zwischen Ringen, Gruppen wie -CF_3 und Gruppen von kleinen Ketten (z.B. Alkylketten) unterschieden wird. Überschreitet die Summe der zuvor berechneten Hydrophobizitätswerte einen bestimmten Mindestwert wird eine hydrophobe Eigenschaft für die Gruppe definiert. Durch die Berücksichtigung der zugänglichen Oberfläche wird der Algorithmus abhängig von der 3D-Konformation des betrachteten Moleküls.[286] Einige Programme (z.B. LigandScout)[299] beschränken sich auf den tabellierten Hydrophobitätswert.

Eine weitere Möglichkeit, auf der viele weitere Ansätze beruhen, ist die regelbasierte Definition von hydrophober Eigenschaft. Dieser Ansatz ist auch für 2D-Methoden geeignet wie CATS2D[175] zeigt. Lipophile Atome werden hierbei SMARTS basiert erkannt (Cl, Br, I, Thioether, C-Atome, die nur zu C-Atomen benachbart sind).[244] ZUCCOTTO definiert Hydrophobizität ebenfalls regelbasiert.[291] Hierbei werden alle 5- und 6-gliedrigen Ringe, tert-Butyl-Gruppen, nicht-aromatischen Halogene und jedes Kohlenstoffatom, das mindestens 2 C-Atome von einem Heteroatom oder jedem bereits als hydrophob klassifizierten C-Atom entfernt ist, als hydrophob definiert. KOSSNER verwendet für seinen pharmakophoren Korrelationsvektor DIP^2 ebenfalls SMARTS-basierte Hydrophobizitäts-Erkennung.[300] Hierbei wird zwischen polaren C-Atomen (ein C-Atom, das mindestens zu einem O-, N- oder F-Atom benachbart ist) und nicht-polaren C-Atomen unterschieden, die zusammen mit Br-, I-, S-Atomen, die Gruppe an hydrophoben Atomen bilden. Sind mindestens 2 Atome dieser Gruppe benachbart, wird eine hydrophobe Eigenschaft definiert.

Weitere noch einfachere Ansätze[301] definieren jedes Atom, das keinen HBA oder HBD darstellt als hydrophob, wobei hierdurch weniger Hydrophobizität als die räumliche

molekulare Gestalt beschrieben wird.[286] KOSSNER berücksichtigt ebenfalls die molekulare Gestalt, indem er bei seinem Ansatz zusätzlich noch sogenannte „lumped hydrophobes“ definiert.[300]

Zusammenfassend lässt sich jedoch feststellen, dass die adäquate Definition hydrophober Eigenschaften sehr schwierig ist. Dies mag ein Grund sein, warum bei vielen RG-Typen auf eine explizite Codierung verzichtet wird.

7.5. Aromatische Eigenschaften

Für die Erkennung aromatischer Eigenschaften ist eine Ringerkennung und nachfolgende Aromatizitätserkennung notwendig.

Die Ringerkennung ist meist SSSR-basiert[289], d.h. es wird die sogenannte „kleinste Menge an kleinsten Ringen“ (Abk. SSSR engl. smallest set of smallest rings) bestimmt. Der SSSR ist definiert als die Menge von Ringen ausgehend von der alle anderen Ringsysteme im molekularen Graphen konstruiert werden können, wobei diese Ringe aus möglichst wenigen Atomen bestehen sollten.[120]

Als aromatisch werden Ringe klassischerweise betrachtet, wenn der Ring planar ist, keine exozyklische Doppelbindung aufweist und die Hückel-Regel, nach der ein aromatisches System $4n+2$ π-Elektronen besitzen muss[302], erfüllt wird.[289, 303] Die Klassifikation eines Ringes als aromatisch ist in einigen Fällen nicht ganz trivial.[303] Je nach zugrundeliegendem Aromatizitätsmodell können Ringsysteme unterschiedlich beurteilt werden.[304]

8. Polypharmakologie und Chemogenomik

8.1. Definitionen und Bedeutung

Polypharmakologie

Polypharmakologie steht im Gegensatz zu dem traditionellen Paradigma „ein Arzneistoff – eine Zielstruktur“[305] und beschreibt das Phänomen, dass ein Molekül Aktivität an mehreren Zielstrukturen zeigen kann.[306] Für die Arzneistoffentwicklung ist dies von zwiespältiger Bedeutung.[307] Die Konsequenzen lassen sich nach PETERS wie folgt zusammenfassen[307]:

1.) Negative Aspekte: Arzneimittelsicherheit/UAWs

Polypharmakologie stellt zum einen ein Sicherheitsproblem dar, da schwerwiegende unerwartete unerwünschte Arzneimittelwirkungen resultieren können. Dies kann schlimmstenfalls, wie bekannte Beispiele in den letzten Jahren gezeigt haben, zu kostspieligen Marktrücknahmen führen. So wurde z.B. das Prokinetikum Cisaprid/Propulsin® aufgrund Arrythmien bedingt durch zusätzliche hERG-Kanal-Blockade oder das Muskelrelaxans Rapacuroniumbromid/Raplon® aufgrund Bronchospasmen bedingt durch zusätzlichen Antagonismus am M_2-Rezeptor zurückgerufen. Da Arzneimittelsicherheit höchste Priorität hat, ist es wichtig, möglichst früh in der Entwicklung entsprechende Risiken zu erkennen, um kostenintensive Weiterentwicklungen zu vermeiden. Daher werden potentielle Wirkstoffkandidaten bereits in den frühen Entwicklungsphasen bezüglich biologischer Aktivität an einer diversen Auswahl an typischen Targets mit sicherheitsrelevantem pharmakologischen Profil („Antitargets“) getestet. Diese in-vitro Testungen sind jedoch teuer und es stehen nur begrenzte Test-Kapazitäten zur Verfügung, sodass eine Vorauswahl relevanter Targets entscheidend ist. Für weitere Details bezüglich Antitargets oder der Erstellung pharmakologischer in-vitro Sicherheitsprofile sei auf ein gleichnamiges Buch[308], Buchkapitel[309] und Übersichtsartikel[310–311] verwiesen.

2.) Positive Aspekte: Verbesserte Wirksamkeit/neue Therapieoptionen

I. Therapie komplexer Erkrankungen

Einige Erkrankungen (z.B. psychische Erkrankungen wie Depression oder Schizophrenie, vgl. Kapitel 1) beruhen auf einem komplexen Zusammenspiel pathophysiologischer Prozesse. Die Modulation mehrerer mit einer Krankheit-assozierten Targets ermöglicht eine komplexe Wirkung, die der Wirksamkeit von selektiven Arzneistoffen deutlich überlegen sein kann. Ein Muster-Beispiel hierfür stellt die Gruppe der Antipsychotika und Antidepressiva dar. Viele der in den letzten Jahren zugelassenen oder in der klinischen Prüfung befindlichen Arzneistoffen aus dieser Gruppe stellen trotz des erhöhten Risikos fürs UAWs aufgrund der besseren Wirksamkeit Multi-Target-Arzneistoffe dar (vgl. Übersicht bei PETERS[307]).

II. Neue Therapieoptionen

a) Vermeidung von Resistenzbildung

Durch Multi-Target-Arzneimittel wird auch die Therapie bisher schwer therapierbarer Erkrankungen (z.B. Multikinase-Inhibitoren bei Tumorerkrankungen

oder Antiinfektiva) ermöglicht. Durch die Modulation von Targets in parallelen Signal-Kaskaden lässt sich auch häufig auftretende Resistenzbildung im Bereich der Tumor- und Infektionskrankheiten reduzieren (vgl. Kapitel 1)

b) Reduktion von UAWs

In einigen Fällen ist es sogar möglich, das Risiko von UAWs durch Arzneistoffe mit polypharmakologischem Profil zu reduzieren, da die Modulation mehrerer krankheitsrelevanter Targets häufig zu einer synergistischen Wirkung ohne Erhöhung der UAWs führt („selektiver Synergismus“). Dies ist der Fall, wenn die UAWs durch die Interaktion mit dem therapeutischen Target (sogenanntes „On-Target“) statt mit therapeutisch-irrelevanten Targets (sogenannte „Off-Targets“) resultieren. Ein Beispiel für einen Multi-Target-Arzneistoff mit reduziertem UAW-Profil stellt das 2010 zugelassene duale Opioid-Analgetikum Tapendalol/Palexia® dar, das zusätzlich zur agonistischen Wirkung am μ-Opioid-Rezeptor die Noradrenalin-Wiederaufnahme hemmt, wodurch die analgetische Wirkung potenziert wird. Aufgrund geringerer μ-Rezeptor-Affinität kann so eine gleiche analgetische Wirkung bei reduziertem μ-Rezeptor-bedingtem UAW-Profil erreicht werden.

c) Verbesserung der Patienten-Compliance/vereinfachte Zulassung

Bei einigen Erkrankungen (z.B. Hypertonie, HIV) ist zur effektiven Therapie eine Kombination mehrerer Arzneistoffe (z.B. Betablocker und ACE-Inhibitor) notwendig. Durch die Entwicklung eines Arzneistoffes, der die Wirkung mehrerer Arzneistoffe vereint, könnte beispielsweise die Einnahme deutlich vereinfacht und eine höhere Therapietreue erreicht werden. Im Vergleich zur Kombination der Einzel-Arzneistoffe in einer Arzneiform wäre hierbei die Zulassung deutlich vereinfacht. Jedoch ist hierbei zu beachten, dass die Optimierung eines Arzneistoffes, der an einem Target angreift, bereits schwierig ist. Die Entwicklung von Multi-Target-Arzneistoffen stellt daher eine große Herausforderung dar. Ähnliche oder überlappende Pharmakophore stellen eine wichtige Voraussetzung für eine erfolgreiche Entwicklung dar, v.a. im Hinblick auf die Vermeidung extremer molekularer Eigenschaften (z.B. hohe Molekularmasse).

III. „Drug Repurposing“

Unter „Drug Repurposing“ wird die Verwendung eines alten Arzneistoffes für ein neues Indikationsgebiet verstanden. Dies folgt dem Prinzip: „Der ertragreichste Weg ein neues Arzneimttel zu finden, ist es mit einem alten zu starten.“ (Medizin-Nobelpreis-Träger 1988 Sir James Black)[312] Die selektive Optimierung von UAWs bzw. einer Off-Target-Aktivität (WERMUTH: „SOSA“-Ansatz, engl. selective optimization of side activities[313]) zur Hauptwirkung (On-Target-Aktivität) stellt eine wichtige Alternative zum HTS in der Entwicklung neuer Arzneistoffe dar. Ein Beispiel hierfür stellen die Sulfonylharnstoffe dar, die aus antiinfektiven Sulfonamiden mit hypoglykämischer UAWs entwickelt wurden. Eine Zusammenstellung einer Vielzahl weiterer Arzneistoffe, die nach diesem Prinzip gefunden wurden, findet sich bei WERMUTH[313] und PETERS[307].

Chemogenomik

Chemogenomik stellt eine interdiziplinäre Schnittstelle von Chemie und Biologie dar und lässt sich als Integration des Target- und Ligandenraums mit dem Ziel der Identifizierung aller Liganden aller Targets beschreiben.[314–315] Chemogenomische Daten lassen sich als 2D-Matrizen mit Zielstrukturen/Genen als Spalten, Molekülen als Zeilen und Bioaktivitätsdaten als numerische Werte beschreiben, wobei diese Matrizes nur sehr spärlich besetzt sind, da Moleküle meist nur an sehr wenigen Targets getestet sind.[315] Basierend auf diesen Informationen versuchen Chemogenomik-basierte Ansätze mittels Ähnlichkeitsanalyse von Liganden oder Targets Beziehungen zwischen Targets, Liganden oder Target-Liganden-Paaren herzustellen.[316] Zwei Grundannahmen sind dafür wichtig[315]: (I) Nach dem in Kapitel 2.3 vorgestelltem SPP weisen ähnliche Liganden ähnliche Eigenschaften auf. Davon ausgehend ist zu erwarten, dass ähnliche Liganden auch an ähnliche Zielstrukturen binden. (II) Im Umkehrschluss sollten Zielstrukturen, die ähnliche Liganden binden, auch ähnliche Bindestellen aufweisen. In Zeiten von Polypharmakologie liefert die Chemogenomik wertvolle Informationen für die Arzneistoffentwicklung.[317–318] Für weitere Details sei auf ein Buch[319], Buchkapitel[317] bzw. eine Vielzahl von Übersichtsartikel[314–315, 318] verwiesen.

Molekül-Promiskuität und „priviligierte Strukturen“

Im Kontext von Polypharmakologie beschreibt *Molekül-Promiskuität* das Verhalten einzelner Moleküle spezifisch an verschiedene Targets zu binden.[320] Ein wichtiges Konzept im Zusammenhang mit Molekül-Promiskuität ist die *„priviligierte Struktur“*[321]. Es handelt sich um molekulare Grundgerüste mit vielseitigen Bindungseigenschaften, die die Bindung an eine Vielzahl verschiedener Targets ermöglichen[322] bzw. gemeinsame Substrukturen von Liganden diverser Targets.[323]

8.2. Bisherige Ansätze

Analyse von Polypharmakologie/Chemogenomik

Obwohl Polypharmakologie erst in den letzten Jahren in den Fokus des Interesses gerückt ist und auch der Zweig der Chemogenomik noch relativ jung ist, sind bereits zahlreiche Chemogenomik-basierte in-silico Ansätze bzw. Ansätze zur Analyse von polypharmakologischer Beziehungen beschrieben. Der nachfolgende Abschnitt soll daher nur einen groben Überblick über wichtige Konzepte geben, darüberhinaus sei auf eine Vielzahl sehr empfehlenswerter Übersichtarbeiten verwiesen. Einen umfassenden Überblick über verschiedenste ligand- und strukturbasierte chemogenomische Methoden gibt beispielsweise ROGNAN[315] in einem hervorragenden Review. Ebenfalls einen guten Überblick über verfügbare in-silico Methoden zur Analyse von Polypharmakologie und „Drug repurposing“ geben ACHENBACHER et al.[324] Ein Review von HU und BAJORATH[320] fasst verschiedende Analysen im Hinblick auf Promiskuität von Molekülen zusammen. Möglichkeiten der Visualisierung von Chemogenomik-Daten sind bei HU et al.[325] zusammengestellt. Verweise auf verschiedene Analysen zur “Off-Target”-Vorhersage finden sich in dem Übersichtartikel von REDDY und ZHANG.[306] Verschiedene Ansätze des „Target Fishing“, d.h. der Vorhersage potentieller „On-Targets“ von Molekülen, fassen JENKINS und BENDER zusammen.[326] Für Target- und UAW-Vorhersagen mittels „biologischer Fingerabdrücke“ oder „Bioaktivitätsspektren“ sei auf BENDER et al.[327] und SCHEIBER et al.[328] verwiesen.

Wie üblich für CADD-Methoden kann zur Unterteilung auch hier grob zwischen ligand- und strukturbasierten Ansätzen unterschieden werden (vgl. ROGNAN[315]). Während *ligandbasierte* Verfahren Liganden-Information (1D-, 2D- oder 3D-Molekül-Repräsentation, vgl. Abschnitt 2.3.2) verwenden und auf Ähnlichkeits-Analyse zwischen Liganden beruhen, verwenden *strukturbasierte* Verfahren Target-Information (Protein-Sequenz, 2D- oder 3D-Protein-Struktur) und beruhen auf der Analyse von Ähnlichkeiten zwischen Targets.[315] Je nach Fragestellung (Ähnlichkeitsanalyse von Zielstrukturen, On- und Off-Target-Vorhersage für Liganden, Vorhersage von UAWs etc.) können dabei verschiedenste Ansätze unterschieden werden. Wie auch im Bereich der SAR-Analyse aufgezeigt (vgl. Abschnitt 2.6.4) sind auch hierbei Netzwerk-basierte Ansätze sehr populär.[329] Ein wichtiger limitierender Faktor aller Ansätze stellt die Unvollständigkeit der verfügbaren Chemogenomik-Daten (spärlich besetzte Ligand-Target-Bioaktivitätsmatrix), der mit einem Bias entsprechender Analysen einhergeht.[330] Bezüglich der allgemeinen Qualität von öffentlich verfügbaren Chemogenomik-Daten sei auf eine gute aktuelle Zusammenfassung von KALLIOKOSKI et al.[96] verwiesen.

Im Folgenden werden nur ausgewählte Aspekte aus dem großen Spektrum verfügbarer Ansätze aufgezeigt und diskutiert, die im Hinblick auf die in dieser Arbeit durchgeführte Analyse von Interesse sind.

1.) Methoden zur Ähnlichkeitsanalyse von Zielstrukturen

Targetbasierte Methoden

Target-basiert wird die Ähnlichkeit zwischen Zielstrukturen üblicherweise durch Sequenz- oder 3D-Struktur-Vergleich bestimmt. Basierend auf den hergestellten Ähnlichkeits-Beziehungen können beispielsweise Off-Target-Beziehungen hergestellt werden (vgl. WEBER et. al.[331]) oder durch Homologie-basierten Ähnlichkeitssuche Liganden für ein neues Target gefunden werden (vgl. SCHUFFENHAUER et al.[246] und FRIMURER et al.[332])

Sequenzbasierte Ähnlichkeit

Da Sequenzlängen selbst innerhalb einer Proteinfamilie stark varrieren können und zusätzlich Deletion/Insertion die Sequenzüberlagerung (z.B. mittels BLAST[333–334]) erschwert, werden meist nur konservierte Motive oder Bindetaschen-Aminosäure-Ketten, die für die Protein-Ligand-Bindung relevant sind und somit Selektivität und Affinität bestimmen, innerhalb einer Proteinfamilie verglichen.[315] Sequenzbasierte Ähnlichkeit kann zur Analyse phylogenetischer Verwandschaftsbeziehungen *innerhalb* einer Proteinfamilie (z.B. SURGAND et al.: GPCRs[335], MANNING: Kinasen[336], RAWLINGS et al. bzw. MEROPS Datenbank: Proteasen[337–338]) verwendet werden.

3D-Strukturbasierte Ähnlichkeit:

Die Voraussetzung für 3D-strukturbasierte Target-Vergleiche ist das Vorhandensein von Kristallstruktur-Information (ggf. von ähnlichen Targets zur Homologiemodellierung).[315] Diese ist für einige Targetfamilien nur in geringem Maß verfügbar, sodass nur sequenzbasierte Vergleiche möglich sind. Der Vorteil von 3D-strukturbasierten Vergleichen ist, die Möglichkeit des Vergleiches von Targets *unterschiedlicher* Proteinfamilien (z.B. Cyclooxygenase-2 und Carbonanhydrase[331] oder KUHN et al.[339]). Üblicherweise wird auch hierbei nur die Bindetasche verglichen.[315] Bindetaschen-Erkennung und die adequate Codierung der Bindetaschen für den Vergleich sind hierbei entscheidend für die Ähnlichkeitsanalyse (vgl. z.B. SCHMITT et al.[340]). Einen Überblick über verschiedene Methoden des Bindetaschenvergleiches geben VULPETTI et al.[341]. Für weitere Verfahren sei auf ROGNAN verwiesen.[315]

Ligandbasierte Methoden

Durch die Verfügbarkeit einer immer weiter zunehmenden Zahl an Bioaktivitätsdaten (vgl. Abschnitt 2.2.2), sind auch immer mehr Liganden für ein Target bekannt. Dadurch stellen ligandbasierte Verfahren eine vielversprechende Alternative zu targetbasierten Ansätzen unter Umgehung der oben genannten Probleme dar (Vergleich verschiedener Targetklassen möglich, keine Kristallstruktur-Information notwendig).

Ligandenvergleich

Die einfachste Möglichkeit des ligandbasierten Vergleichs von Zielstrukturen zur Erstellung pharmakologischer Netzwerke stellt die Bestimmung gemeinsamer Liganden und ein Vergleich ihrer Eigenschaften dar. PAOLINI et al.[342] nutzen dieses Prinzip für ihre Analyse, wobei sie zusätzlich zur Charakterisierung der Stärke der pharmakologischen Target-Interaktion die Bioaktivitätsdifferenzen der Liganden der beiden Targets berücksichtigten.

Substruktur-basierte Ähnlichkeit

Für GPCRs zeigten VAN DER HORST et al.[343], dass basierend auf der Bestimmung häufiger Substrukturen („frequent substructure mining“) und der Korrelation der Häufigkeiten der häufigsten Substrukturen aller Targets ebenfalls ligandbasierte Target-Beziehungen hergestellt werden können. Ebenfalls substrukturbasiert mittels Molekül-Fragmentierung und anschließender Bestimmung der Fragment-Ähnlichkeit der Targets (Tc-Ähnlichkeit auf Basis der häufigsten Fragmente) stellten SUTHERLAND et al.[344] ligandbasierte Targetbeziehungen zwischen Kinasen, aber auch anderen Target-Klassen, her. Wie zu erwarten, hat sich bei beiden Analysen im Vergleich zu sequenzbasierter Ähnlichkeit gezeigt, dass wie erwartet oftmals hohe ligandbasierte Ähnlichkeit zwischen Targets mit hoher Sequenzähnlichkeit gefunden wird, jedoch auch Targets mit niedriger Sequenzähnlichkeit aufgrund ähnlicher Liganden verknüpft werden (potentielle Off-Targets).

Fingerprint-basierte Ähnlichkeit

Ein wichtiger auf Fingerprint-Ähnlichkeit (u.a. Daylight-FP, ECFP4, CATS, FEPOPS[345–346]) beruhender Ansatz zur Erstellung von ligandbasierten Target-Netzwerk-Beziehungen, der die Grundlage für eine Vielzahl weiterer Analysen[346–349] darstellt, ist der von KEISER et al. entwickelte *Similarity Ensemble Approach* (Abk.: SEA)[345]. SEA verwendet statistische Techniken, die dem BLAST-basierten Sequenzvergleich zu Grunde liegen, zur Bestimmung der Ähnlichkeit von Ligandensätzen. Hierfür wird ein Rohwert basierend auf der Summe aller Ligandenpaare zweier Targets bestimmt, der stark von der Datensatzgröße abhängig ist. Durch Bestimmung des Datensatzgrößen-unabhängigen Z-Wertes (basierend auf Zufalls-Ligandensätzen) und basierend auf einer Z-Wert-Verteilung (Berücksichtigung verschiedener Ähnlichkeits-Schwellenwerte) wird ein finaler E(rwartungs)-Wert bestimmt, der die Wahrscheinlichkeit für einen bestimmten Z-Wert bei Verwendung von Zufalls-Daten angibt. Umso kleiner ein E-Wert für ein Target-Paar ist, desto wahrscheinlicher ist eine Ähnlichkeitsbeziehung.

Basierend auf dem Atom-zentrierten pharmakophoren Eigenschafts-Fingerprint SHED (SHannon Entropy Descriptor)[350] haben GREGORI-PUIGJANÉ und MESTRES[351] Beziehungen verschiedenster Targetklassen bzw. innerhalb der Familie der nukleären Rezeptoren (MESTRES et al.[352]) analysiert und erfolgreich off-Targets vorhergesagt.

2.) Methoden zur (Off-)Target-Vorhersage

Der Begriff „Target Fishing“ umschreibt anschaulich die Vorhersage von potentiellen (On-/Off-)Targets für ein bestimmtes Molekül von Interesse.[326] Verschiedene Ansätze wurden bereits erfolgreich zur Vorhersage von (Off-)Targets entwickelt, wobei entsprechende Vorhersagen entweder mittels Literaturrecherche[348] oder experimenteller Testung[347–348] validiert wurden.

Ligandbasiert sind zahlreiche Fingerprint-basierte Ansätze (u.a. unter Verwendung des SEA-Prinzipes[347–348] oder Bayesian Klassifizierer[353–355]) beschrieben. Aber auch ligandbasierte Ansätze unter Verwendung der 3D-Molekülstruktur (z.B. Vergleich der Überlappung der als Gauß-Peaks repräsentierten Molekülvolumina[356–357] mittels Rapid Overlay of Chemical Structures (Abk. ROCS)[358–359] oder Pharmakophor-basiert wie z.B. PharmMapper[360] oder auf Basis der FEature POint PharmacophorS[361–362]) sind beschrieben.

Eine wichtige strukturbasierte Methode stellt das „inverse Docking“ (z.B. TarFisDock[363], INVDOCK[364]) dar. Hierbei werden nicht wie normalerweise mehrere Liganden in die Bindetasche eines Targets eingepasst und resultierende Bindungsmodi bewertet, sondern ein Molekül wird in mehrere Targets gedockt.[315] Die aus der verwendeten Bewertungsfunktion resultierenden Docking-Scores für die besten Docking-Posen werden dann zum Target-Ranking verwendet. Limitiert ist dieser Ansatz abermals auf Zielstrukturen für die Kristallstrukturen oder Homologiemodelle verfügbar sind, sowie durch das Fehlen optimaler Bewertungsfunktionen[365].

3.) Methoden zur Vorhersage von UAWs

Die Vorhersage schwerer unerwünschter Arzneimittelwirkungen ist wichtig in der präklinischen Sicherheitsheitsprofil-Erstellung neuer Arzneistoffkandidaten.[328] Verfügbare Ansätze vereinen Methoden des Text-Mining in entsprechenden Dankenbanken in Kombination mit vorgenannten chemogenomischen Techniken.[366] Eine netzwerkbasierte Verknüpfung von UAWs und FP-basierter Liganden-Ähnlichkeit[367] ist hierbei ebenso möglich wie eine UAW-basierte Analyse von Target-Ähnlichkeiten[368].

9. Zielsetzung der Arbeit

Das primäre Ziel dieser Arbeit war die Entwicklung einer Methode zur Visualisierung und Analyse von Struktur-Aktivitäts-Beziehungen in großen Datensätzen. Hierbei sollte zum einen im Vordergrund stehen, dass die zu entwickelnde Methode auf einem Konzept beruht, das für den medizinischen Chemiker intuitiv zu verstehen ist und leichter interpretierbar ist als beispielsweise auf Fingerprint-Ähnlichkeit-basierende Verfahren. Hierfür erschien das Konzept der *maximal gemeinsamen Substruktur (MCS)* eine vielversprechende Grundlage (vgl. Abschnitt 2.6.5 und Kapitel 4)

Netzwerk- oder Graphen-basierte Ansätze haben sich als sehr gut geeignet erwiesen, große Datenmengen und komplexe Sachverhalte kompakt zusammenzufassen und zu veranschaulichen (vgl. Abschnitt 2.6.4). Wie bei der Scaffold-basierten SAR-Analyse beschrieben (vgl. Abschnitt 2.6.3), sind zudem chemische Beziehungen sehr intuitiv durch eine hierarchische Darstellung zu erfassen. Aus diesen Gründen sollte der zu entwickelnden Methode eine *hierarchische Netzwerk-Struktur* zu Grunde gelegt werden.

Wie in Kapitel 2.6 bereits aufgezeigt, ist ein weiteres Problem von vielen verfügbaren Verfahren, dass sie oftmals nur zur Analyse von Analogserien in Datensätzen geeignet sind. Zudem ist, wie in Kapitel 4.1 dargelegt, das Konzept des MCS aufgrund der Komplexität der zugrunde liegenden Berechnungen wenig zur Anwendung auf große Datensätze geeignet, die beispielsweise mehr als tausend Moleküle umfassen. Zudem sind die MCSs aufgrund der Notwendigkeit der exakten Molekülpaarung oftmals sehr klein. Um die genannten zu lösen, erschien in Anlehnung an GARDINER et al.[232] die *Kombination mit reduzierten Graphen* (RGs) ein erfolgversprechender Lösungsansatz (vgl. Abschnitt 5.5) für die Analyse von großen Datensätzen. Wie in Abschnitt 2.2.1 dargestellt, ist biologische Aktivität weniger an eine bestimmte Substruktur gebunden als an bestimmte pharmakophore Eigenschaften. Die Abstraktion des Moleküls auf Merkmale, die potentiell zu einer Protein-Ligand-Interaktion führen können, stellt somit einen alternativen, vielversprechenden Ansatz der SAR-Analyse (im Vergleich zu den in Kapitel 2.6 vorgestellten Methoden) dar. Das Aufzeigen von gemeinsamen pharmakophoren Eigenschaften ist ein wichtiger Vorteil, der von hohem Wert für die SAR-Interpretation sein kann, inbesondere der SAR-Analyse von heterogeneren Datensätzen.

Unter Berücksichtigung der genannten Voraussetzungen ist die inSARa Methode entstanden, die die Vorteile von Substruktur- und RG-basierter SAR-Analyse (vgl. Abschnitt 2.6.5 und 2.6.6) durch Verwendung der synergistischen Kombination von MCS und RG vereint. Ihr Grundprinzip ist eine *hierarchische Netzwerk-Struktur basierend auf klar-definierten Substruktur-Beziehungen von gemeinsamen pharmakophoren Eigenschaften (RG-MCSs)*.

Ziel 1: Implementierung der inSARa-Methode

Das erste Ziel dieser Arbeit war die Implementierung der oben charakterisierten inSARa-Methode. Hierbei wurden folgende methodische Schwerpunkte gelegt:

a) *Implementierung der Codierung der Datensatz-Moleküle als reduzierte Graphen*: Hierzu wurde versucht in der Literatur beschriebene Ansätze (vgl. Kapitel 5.3)

nachzuimplementieren bzw. weiterzuentwickeln. Dabei wurde analysiert, welches Abstraktionsniveau zur Analyse von SARs (unter Berücksichtigung der in Kapitel 6 und 7 zusammengefassten Informationen) am besten geeignet ist. Nachfolgend wird nur die Implementierung beschrieben, die sich als vorteilhaft für eine große Zahl verschiedener Zielstrukturen erwiesen hat.

b) *Algorithmus zur Erzeugung einer hierarchischen MCS-basierten Netzwerk-Struktur*.

c) Realisierung einer einfachen und intuitiven *Visualisierung der inSARa-Netzwerke* für die interaktive SAR-Interpretation.

Ziel 2: Optimierung und Vergleich der inSARa-Methode

Das zweite Ziel dieser Arbeit war die Optimierung der inSARa-Methode für die Analyse großer Datensätze und ein Vergleich des RG-MCS-Prinzipes mit der häufig für SAR-Analysen verwendeten Fingerprint-Ähnlichkeit. Folgende Arbeits-Schwerpunkte wurden hierbei definiert:

a) Analyse von *unspezifischer Ähnlichkeit in Zufallsmolekülpaaren*

b) Analyse des *Einflusses von verschiedener Faktoren auf Netzwerk-Topologie/Komplexität*

c) *Vergleich mit Fingerprint-basierter Ähnlichkeit*

Ziel 3: Anwendung der inSARa-Methode

Das dritte Ziel der Arbeit war die Anwendung der optimierten inSARa-Methode zur Überprüfung der Leistungsfähigkeit der Methode im Vergleich mit anderen Verfahren. Hierbei wurden folgende Anwendungsschwerpunkte untersucht:

a) *Interaktive SAR-Interpretation* an einem einzelnen Target.

b) *Automatisierte SAR-Analyse* eines Targets bzw. zum Vergleich verschiedener Targets

c) *Ligand-basierte Analyse von Target-Ähnlichkeiten und Kreuzreaktivitäten* (in Anlehnung an Ansätze zur Analyse von Polypharmakologie bzw. Chemogenomik-Methoden aus Abschnitt 8.2)

II. Methoden

10. Die inSARa-Methode

10.1. Überblick: Das Prinzip

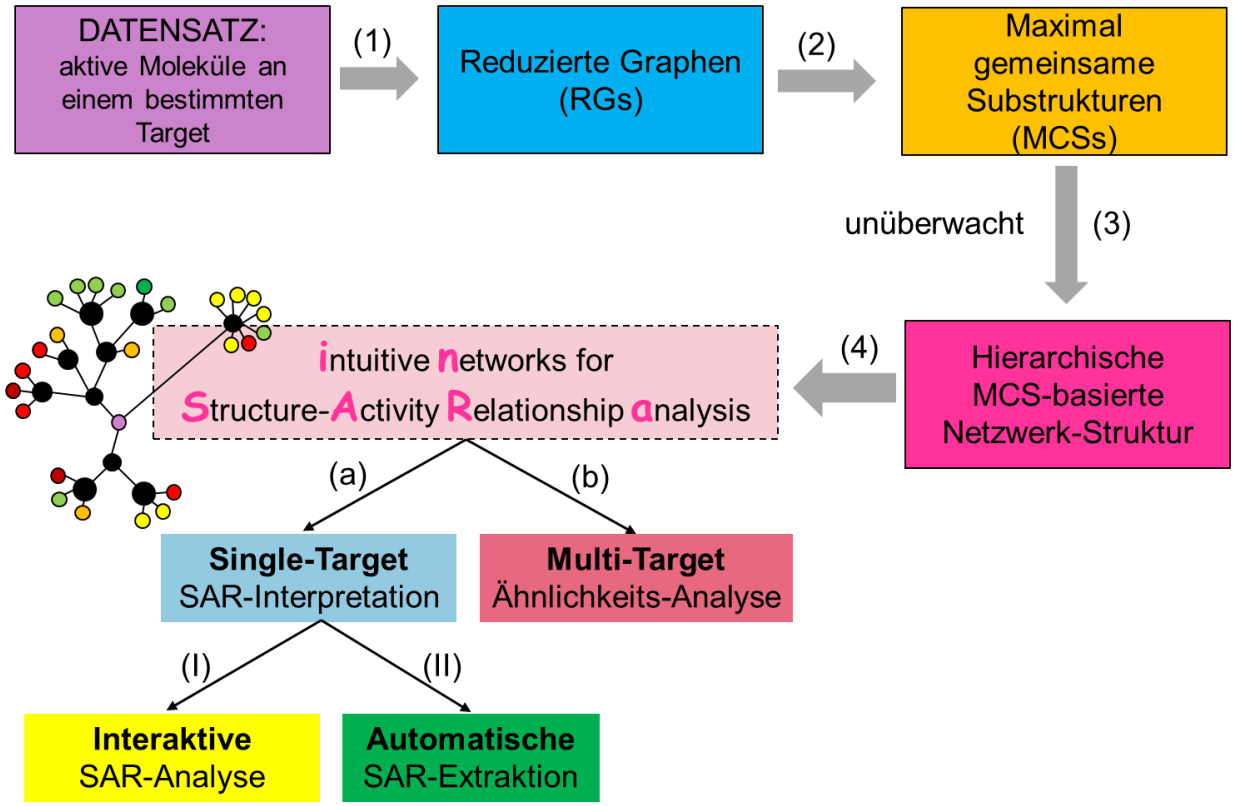


Abbildung 10.1. Überblick über das Prinzip der inSARa-Methode.

Das Fließschema in Abbildung 10.1 fasst das Prinzip der inSARa-Methode zusammen. Voraussetzung für inSARa ist ein Datensatz bestehend aus an einer bestimmten Zielstruktur getesteten, biologisch aktiven Molekülen. Diese Moleküle sollten wie in Kapitel 12 beschrieben standardisiert und vorbereitet werden. Der erste wichtige Schritt (1) ist dann die Umwandlung dieser Datensatzmoleküle in reduzierte Graphen mittels SMARTS-basierter Erkennung von pharmakophoren Eigenschaften (vgl. Abschnitt 10.2). Als Nächstes (2) wird dann paarweise zwischen allen RGs die maximal gemeinsame Substruktur bestimmt (vgl. Abschnitt 10.3). Aus dieser Gesamt-Menge an MCSs wird dann (3) unüberwacht, d.h. ohne Berücksichtigung von Bioaktivitätsinformation, eine hierarchische Netzwerk-Struktur aufgebaut, indem kleinere MCSs jeweils mit den nächst-größeren verbunden werden (vgl. Abschnitt 10.4). Nach einigen Prozessierungs-Schritten zur Reduzierung der Netzwerk-Komplexität werden die Datensatzmoleküle den passenden MCSs-Knoten zugeordnet (4) und somit die finalen inSARa-Netzwerke erhalten. Die Visualisierung und Aufbau der inSARa-Netzwerke wird in den Abschnitten 10.5 und 10.6 erläutert. Die resultierenden

Netzwerke bieten vielfältige Anwendungsmöglichkeiten, die in späteren Kapiteln im Detail vorgestellt und diskutiert werden.

In den nachfolgenden Abschnitten findet sich zunächst eine detaillierte Beschreibung der Grundmethode, gefolgt von einigen Analysen zur Optimierung und Charakterisierung der Netzwerke.

10.2. Schritt 1: Umwandlung der Moleküle in Reduzierte Graphen

Der erste wichtige Schritt in der inSARa-Methode ist die Umwandlung der Datensatz-Moleküle in reduzierte Graphen. Verschiedene Molekül-Codierungs-Schemata wurden in Anlehnung an die in Kapitel 5.3 beschriebenen RG-Typen in Voranalysen getestet. Der im Folgenden beschriebene Ansatz hat sich als derjenige mit den besten Gesamtergebnissen bezogen auf die damit analysierten Datensätze herausgestellt. Nichtsdestotrotz ist es wichtig zu betonen, dass inSARa auch mit anderen RG-Typen kombiniert werden kann. Die einzige Voraussetzung hierfür ist, dass die vom Benutzer bereitgestellten RGs als Graphen repräsentiert werden. Zudem ist es auch bei diesem RG-Umwandlungs-Schema möglich, die RGs durch Anpassen der SMARTS Definitionen (vgl. Abschnitt 10.2.3), die für die Eigenschafts-Erkennung verwendet werden, zu variieren. Diese Flexibilität ist insofern wichtig, als dass die erfolgreiche Erfassung von Ähnlichkeit zwischen den aktiven Molekülen, wie in Kapitel 2.3 betont, stark abhängig von der passenden RG-Definition ist. Es wird nicht möglich sein eine universell optimale Definition zu finden. Zur Optimierung einer SAR-Analyse ist es aus diesem Grund immer empfehlenswert, die RG-Definition an die Besonderheiten der zu analysierenden Zielstruktur und des Datensatzes anzupassen.

10.2.1. Vergleich der RG-Implementierung mit bisherigen Ansätzen

Im Folgenden werden die Besonderheiten des verwendeten RG-Typen hervorgehoben und mit bisherigen Ansätzen, die in Kapitel 5.3 beschrieben wurden, verglichen.

Die an der Universität von Sheffield entwickelten RGs dienten als Grundlage für die für inSARa optimierten RGs. Die in dieser Arbeit verwendete RG-Implementierung basiert auf einer modifizierten Variante der Ar/F(4) RG-Definition von GILLET et al.[235], die von BARKER et al. um negative ionisierbare (saure) und positiv ionisierbare (basische) Eigenschafts-Knoten erweitert wurde[240]. Dieses RG-Grundschema wurde auch von GARDINER et al.[232] für ihre MCS-basierten Analysen verwendet. Im Gegensatz zu HARPER et al.[201] und BIRCHALL et al.[216], werden PI- und NI-Eigenschaften jedoch nicht mit Strukturtyp-Annotationen (azyklisch, aliphatischer Ring, aromatischer Ring) kombiniert. Das bedeutet, dass der gesamte Ring als NI oder PI codiert wird, sofern eine saure oder basische Eigenschaft oder negative oder positive Ladung in diesem Ring gefunden wird. Ein Beispiel hierfür findet sich in Abbildung 10.2b, wo die Amidin-Struktur partiell in den aliphatischen 6-Ring eingebunden ist. Folglich kann man bei NI- oder PI-Knoten aus dem entsprechenden Pseudoatom nicht schließen, ob es sich um einen zyklischen oder azyklischen Molekülteil gehandelt hat. Diese stärkere Abstraktion wurde gewählt, um zu

verhindern, dass der RG aus zu vielen bzw. zu spezifischen Knotentypen besteht und so bestimmte Ähnlichkeiten nicht mehr erkannt werden.

Bei diesem Ansatz wird der von HARPER et al.[201] publizierte Pseudoatom-Code verwendet und entsprechend für PI- und NI-Eigenschaft aus den oben genannten Gründen angepasst. Das bedeutet, dass alle PI-Eigenschaften durch Nb und alle NI-Eigenschaften durch Mo codiert werden (vgl. Tabelle 10.1). Im Original-Schema von HARPER et al. codierten diese Übergangsmetalle nur für NI-/PI-Eigenschaften in azyklischen Molekülteilen.

Linker werden in diesem RG-Schema ebenfalls explizit codiert. Sie sind definiert als alle nicht-terminalen azyklischen Atome ohne NI-/PI- oder HBA-/HBD-Eigenschaft. Detaillierte Analysen von BARKER et al. haben gezeigt, dass RGs, die terminal nicht-interagierende Gruppen explizit codieren, gesamt betrachtet am besten abschneiden[240]. Daher werden diese Gruppen (z.B. terminale Alkylgruppen oder Halogen-haltige Gruppen) hier ebenfalls bei der Codierung berücksichtigt. Es wird das gleiche Pseudoatom (Zn), das auch gleichzeitig Linker-Atome repräsentiert, verwendet. Bei HARPER et al.[201] (und auch BIRCHALL et al.[216]) wurden terminale Gruppen nicht weiterberücksichtigt, Zn codiert hier nur für Linker-Atome. Diese Doppelcodierung durch ein Pseudoatom kann zu Fehlpaarungen bei der MCS-Bestimmung führen. Es ermöglicht jedoch auch einen höheren Abstraktionsgrad und reduziert sogleich die Anzahl an RG-Knoten-Typen.

Besitzt eine funktionelle Gruppe sowohl Donor- als auch Akzeptor-Eigenschaften wird ihm eine gemeinsame Donor-Akzeptor-Eigenschaft (Abk. HBAD) zugeordnet. BARKER et al. folgend werden zudem annelierte Ringsysteme als mehrere RG-Knoten codiert.[240] Annelierung wird analog durch eine Doppelbindung im RG kenntlich gemacht. Vorversuche mit einer Implementierung in Anlehnung an BIRCHALL et al.[216], wo das komplette Ringsystem durch ein Pseudoatom codiert wurde, haben sich für die SAR-Analyse als zu wenig spezifisch herausgestellt.

Alle erlaubten Eigenschaftskombinationen sind Tabelle 10.1 zu entnehmen, die eine komplette Auflistung aller möglichen RG-Knoten-Typen darstellt.

Tabelle 10.1: Pseudoatom-Code für die RGs und die jeweils codierten pharmakophoren Eigenschaften; verändert nach HARPER et al.[201]

Eigenschafts-Definition	**Pseudoatom im RG**
– positiv ionisierbar (PI)	Nb
– negativ ionisierbar (NI)	Mo
(1) **Ring**	
a. **aromatisch**	
– H-Brücken-Akzeptor (HBA)	V
– H-Brücken-Donor (HBD)	Ti
– H-Brücken-Akzeptor und -Donor (HBAD)	Cr
– ohne weitere Eigenschaft	Sc
b. **aliphatisch**	
– H-Brücken-Akzeptor (HBA)	W
– H-Brücken-Donor (HBD)	Ta
– H-Brücken-Akzeptor und -Donor (HBAD)	Re
– ohne weitere Eigenschaft	Hf
(2) **Azyklisch**	
– H-Brücken-Akzeptor (HBA)	Ni
– H-Brücken-Donor (HBD)	Co
– H-Brücken-Akzeptor und -Donor (HBAD)	Cu
– ohne Eigenschaft (Linker oder terminale Gruppe)	Zn

10.2.2. SMARTS-basierte Definition pharmakophorer Eigenschaften

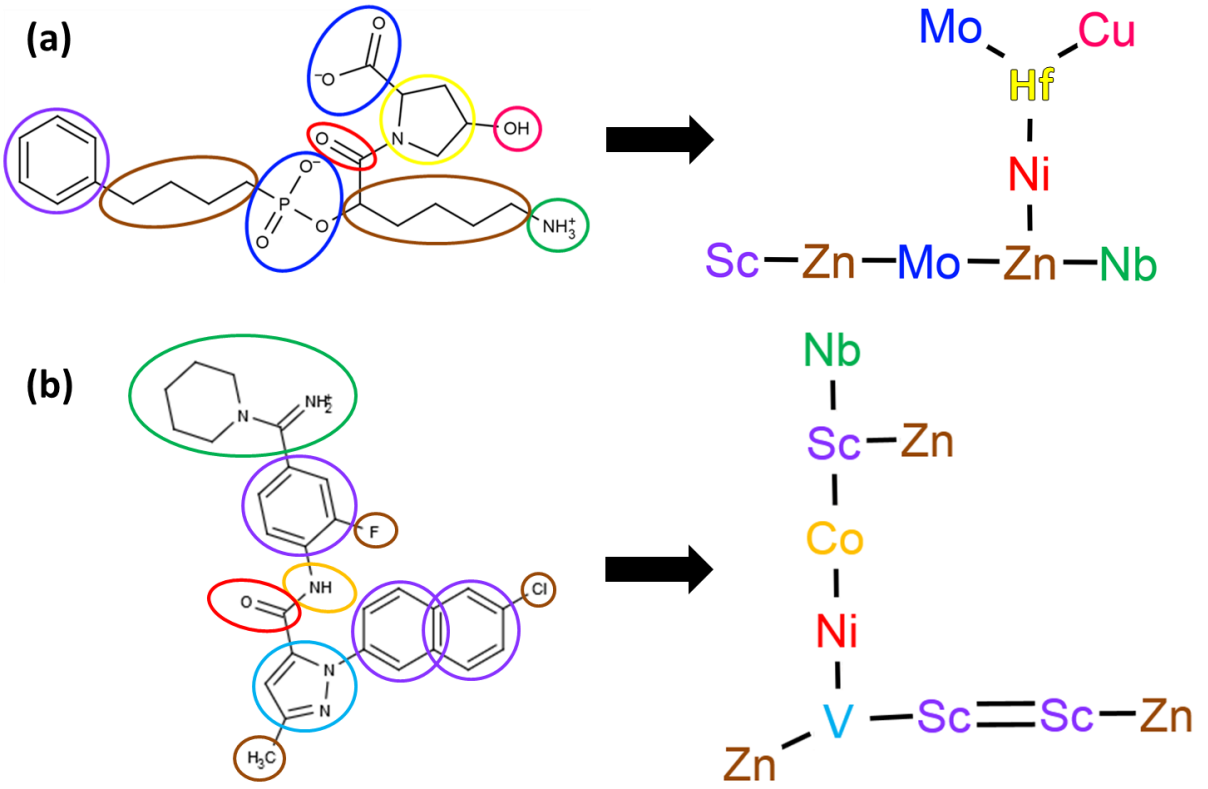


Abbildung 10.2. Umwandlungsprozess in reduzierte Graphen veranschaulicht am Beispiel von zwei verschiedenen Molekülen (a) und (b). Auf der linken Seite der Pfeile sind jeweils die Datensatzmoleküle abgebildet, auf der rechten Seite die resultierenden RGs. Die Definitionen der RG-Knoten finden sich in Tabelle 10.1. Die codierten molekularen Merkmale sind entsprechend der Farbe des zugehörigen Pseudoatoms im RG markiert.

Analog zu den in Kapitel 5.3 beschriebenen RG-Ansätzen werden auch hier bei der Umwandlung eines Moleküls in einen reduzierten Graphen sogenannte SMARTS Ausdrücke zum Auffinden der verschiedenen pharmakophoren Eigenschaften verwendet.

SMARTS Ausdrücke (SMiles ARbitrary Target Specification) stellen eine Erweiterung der in Kapitel 2.3.2 vorgestellten SMILES Notation dar. Die SMARTS Notation ermöglicht die Spezifikation von generischen molekularen Substrukturen und Mustern. Bestimmte Atom- und Bindungseigenschaften (z.B. aromatisch, aliphatisch, im Ring der Größe X) können definiert werden und mit Hilfe von logischen Operatoren (z.B. und/oder/nicht) zu abstrakten Suchanfragen verknüpft werden. Zudem besteht durch die Verwendung von rekursiven SMARTS die Möglichkeit die chemische Umgebung eines Atoms oder einer Bindung zu definieren. Alle SMILES Strings sind auch als SMARTS Ausdrücke verwendbar.[113, 369]

Da die SMARTS Suchen mittels OEChem TK realisiert werden, sind die verwendeten Suchanfragen speziell an die SMARTS Semantik von OpenEye angepasst, die sich in bestimmter Hinsicht von der ursprünglich von Daylight Chemical Information Systems[369] entwickelten Semantik abweicht (z.B. verändert sich aufgrund der fehlenden Implementierung des SSSR-Algorithmus im OEChem TK die Bedeutung von R<n>)[370].

Tabelle 10.2. Molekulare Eigenschaften und funktionelle Gruppen, die durch die SMARTS Ausdrücke bei der Erkennung von pharmakophoren Eigenschaften bei der RG-Erzeugung codiert werden; zusammengestellt nach GILLET et al[235], GREENE et al.[287] und TAMINAU et al.[289] und ZUCCOTTO[291].

Negativ ionisierbares Zentrum (NI)	**Positiv ionisierbares Zentrum (PI)**	**H-Brücken-Akzeptor (HBA)**	**H-Brücken-Donor (HBD)**
• Carbonsäuren • S-/P Säuren (Sulfon-, Sulfin-, Phosphor-, Phosphinsäuren) • Tetrazole • Saure Imide • Saure Sulfonamide • Trifluormethyl-sulfonamide • weitere funktionelle Gruppen, die bei physiologischem pH (7.4) mit hoher Wahrscheinlichkeit deprotoniert vorliegen (gemäß Implementierung in MOE's sdwash) • negative Ladung darf nicht direkt benachbart zu positive Ladung sein (Ausschluss von Sauerstoff-Atomen von Nitrogruppen)	• aliphatische basische Amine • Amidine • Guanidine • weitere funktionelle Gruppen, die bei physiologischem pH (7.4) mit hoher Wahrscheinlichkeit protoniert vorliegen (gemäß Implementierung in MOE's sdwash) • positive Ladung darf nicht direkt benachbart zu negativer Ladung sein (Ausschluss von Stickstoff-Atomen von Nitrogruppen)	• jedes Stickstoff-, Sauerstoff- und Schwefel-Atom mit mindestens einem nicht-delokalisiertem, freiem Elektronenpaar ohne positive Formalladung • Ausschluss von (Sulfon-)Amid-N-Atomen aufgrund von Delokalisierung • Ausschluss von Amididin/ Guanidin-N-Atomen aufgrund von positiver Ladung und Delokalisierung	• Jedes Stickstoff-, Sauerstoff- und Schwefel-Atom mit mindestens einem kovalent gebundenem Wasserstoff-Atom ohne negative Formalladung

Die molekularen Eigenschaften oder funktionellen Gruppen, die jeweils durch die SMARTS Ausdrücke erfasst und als pharmakophore Eigenschaft codiert werden; sind in Tabelle 10.2 zusammengefasst. Die Definitionen für die pharmakophoren Eigenschaften sind unter Berücksichtigung der in Kapitel 7 beschriebenen Überlegungen adaptiert nach GILLET et al.[235], GREENE et al.[287] TAMINAU et al.[289] und ZUCCOTTO[291]. Im Folgenden werden kurz die inSARa zugrunde liegenden Definitionen erläutert.

PI/NI:

inSARa erkennt entsprechende typische funktionelle Gruppen (vgl. Kapitel 7) über SMARTS-Abgleich. Die verwendeten SMARTS versuchen die Delokalisierung der Ladung zu berücksichtigen, sodass alle Atome der funktionellen Gruppe als NI oder PI erkannt werden und entsprechend später zu einem Mo- oder Nb-Pseudoatom umgewandelt werden.

Schwächere Säuren oder Basen oder weitere starke Säuren und Basen, die nicht in der SMARTS-Liste berücksichtigt sind, werden zusätzlich in Abhängigkeit von der Protonierung/Deprotonierung durch MOE's sdwash-Funktion als PI oder NI erkannt.

HBA:

Der Ausschluss von (Sulfon-)Amid-N und Guanidin/Amidin-N aufgrund von Delokalisierung des freien Elektronenpaars wurde von TAMINAU et al.[289] übernommen. Bei Anilin-N hat es sich in den SAR-Analysen als vorteilhaft erwiesen ihn als HBA zu erkennen. Auf die Bestimmung der sterischen Zugänglichkeit bei der HBA-Erkennung wird verzichtet, da hierfür eine dreidimensionale Molekül-Struktur notwendig wäre, deren Erzeugung im Rahmen der inSARa-Methode nicht vorgesehen ist. Aus Gründen der Einfachheit wird ebenfalls auf Ausnahmen wie z.B. den Ausschluss schwacher HBAs verzichtet. Mittels Erweiterung der verwendeten SMARTS-Ausdrücke können diese Ausnahmen jedoch leicht ergänzt werden.

HBD:

Eine Kombination der Definition von GREENE et al.[287] und TAMINAU et al.[289] wird verwendet. In Anlehnung an ZUCCOTTO[291] und GILLET et al.[235] werden Thiole auch als HBD berücksichtigt. Weitere schwache Donoren (wie z.B. Acetylen-Gruppen) werden analog zu den vorgenannten Ansätzen aus Gründen der Einfachheit nicht als HBD definiert. Auch hier ist eine Erweiterung der SMARTS-Liste zur Berücksichtigung dieser Spezialfälle möglich.

Hydrophobe Eigenschaften:

Aufgrund der in Abschnitt 7.4 beschriebenen Schwierigkeiten bei der korrekten Erkennung, wird in der hier beschriebenen Implementierung auf die explizite Codierung von hydrophoben Eigenschaften (wie auch in den unter 5.3.1 vorgestellten RG-Definition) verzichtet. In Vorversuchen wurden Implementierungen ausprobiert, bei denen Atome einer endständigen Alkylgruppen iterativ zu einem Pseudoatom, das hydrophobe Eigenschaften codiert, zusammengefasst, sowie halogenhaltige Substituenten SMARTS-basiert als hydrophob codiert wurden. Hierdurch werden die RGs spezifischer. Ein Vorteil für die SAR-Analyse konnte bei den analysierten Datensätzen jedoch nicht festgestellt werden. Daher werden anderen Ansätzen folgend nur terminale Gruppen im Allgemeinen codiert, wodurch die Anzahl an RG-Atom-Typen reduziert werden kann.

10.2.3. SMARTS-basierte RG-Umwandlung

Das RG-Umwandlungs-Schema, das im Folgenden beschrieben wird, ist in Abbildung 10.2 veranschaulicht. Die gesamte RG-Erzeugung wurde in Python implementiert unter Verwendung von Routinen des kommerziell erhältlichen OEChem TK[370] und der freiverfügbaren Programmierbibliotheken von Open Babel[114–115]. Der komplette Quellcode findet sich in Abschnitt 27.1 im Anhang.

1.) SMARTS-Suchen (PI-, NI-, HBA-, HBD-Eigenschafts-Erkennung)

Für jedes Molekül wird zu Beginn eine leere Eigenschaftsmatrix angelegt, die im Folgenden mit Information gefüllt wird und als Grundlage für die finale RG-Umwandlung dient. Für jedes Heteroatom des Moleküls wird hier gespeichert, ob bestimmte Eigenschaften erfüllt werden.

Zunächst werden nacheinander die vorhandenen PI-/NI-/HBA-/HBD-Eigenschaften im Molekül mittels zugehöriger SMARTS Ausdrücke (siehe oben) bestimmt. Erfüllen Atome die abgeprüften Eigenschaften, wird dies in der Eigenschaftsmatrix gespeichert. Weist ein Atom sowohl HBA- als auch HBD-Eigenschaften auf, so wird die kombinierte HBDA-Eigenschaft ebenfalls als erfüllt gespeichert.

2.) Ring- und Aromatizitäts-Erkennung

Im nächsten Schritt werden zyklische und azyklische Molekülteile unterschieden. Zyklische Teile werden wiederum zusätzlich in aliphatische und aromatische Bereiche unterteilt. Hierfür ist zunächst eine Ring-Erkennung und anschließend eine Aromatizitäts-Erkennung notwendig. In der Eigenschaftsmatrix wird jeweils gespeichert, ob ein Atom Teil eines Ringes ist und ob dieser als aromatisch klassifiziert wird.

Für die Ring-Erkennung wird zunächst die „kleinste Menge an kleinsten Ringen“ (vgl. Abschnitt 7.5) unter Verwendung von dem auf Breitensuche-basierenden SSSR-Algorithmus von FIGUERAS[371] bestimmt. Da die Bestimmung des SSSR in bestimmten Fällen[372–373] (z.B. bei Molekülen mit besonderer Symmetrie oder Ringkomplexität) nicht deterministisch ist, verwendet OpenEye für seine Ring-Definition nicht den SSSR[304]. Daher wird für die Bestimmung des SSSR Open Babel verwendet.

Es werden nur für diejenigen Moleküle RGs generiert, deren kleinste Ringe aus maximal sieben Ring-Atomen (adaptiert nach STIEFL et al.[239]) bestehen. Makrozyklische Moleküle, die i.d.R. nur einen sehr kleinen Anteil in den analysierten Datensätzen ausmachen, werden nicht umgewandelt und somit von weiteren SAR-Analysen ausgeschlossen.

Im nächsten Schritt werden die erkannten Ringe dann als „aromatisch“ oder „aliphatisch“ klassifiziert. Für die Aromatizitäts-Erkennung wird das im OEChem TK implementierte Standard-Aromatizitäts-Modell „OpenEye“ verwendet.[304] Auch bei diesem Modell, das ähnlich dem „Daylight-Aromatizitäts-Modell“[113] ist, wird ein Ringsystem als aromatisch betrachtet, sofern die Hückel-Regel erfüllt wird (vgl. Abschnitt 7.5). Im Gegensatz zur klassischen Aromatizitäts-Definition werden jedoch auch Moleküle mit exozyklischen

Doppelbindungen wie z.B. das 1H-Pyridin-4-on als „aromatisch“ angesehen.[304] Dies hat den Vorteil, dass Ähnlichkeiten zwischen ähnlichen Ringsystemen leichter erkannt werden. Wenn man die klassische Aromatizitäts-Definition für die RG-Konvertierung verwenden möchte, kann man optional auch das ebenfalls im OEChem TK implementierte „MMFF-Modell“[374] verwenden.

Da im RG Anellierungen zwischen Ringsystemen später durch Doppelbindungen codiert werden, werden Anellierungsstellen, d.h. Atome, die Bestandteil beider Ringsysteme sind, ebenfalls mittels SMARTS bestimmt und gespeichert.

3.) Eigenschafts-Umwandlung

Basierend auf der Eigenschaftsmatrix und zusätzlich gespeicherter Information wird das Molekül nun schrittweise in einen Pseudoatom-basierten Graphen umgewandelt. Bei der Umwandlung wird die von HARPER et al. vorgeschlagene Prioritätsreihenfolge verwendet[201]: PI wird gegenüber NI bevorzugt, wohingegen NI gegenüber den H-Brücken-Eigenschaften (HBA, HBD, HBAD) priorisiert wird. Dies ermöglicht eine einheitliche Umwandlung auch in den Fällen, wo einem Atom mehr als eine pharmakophore Eigenschaft zugeordnet werden kann. So wird beispielsweise eine Carboxylat-Gruppe als NI codiert, obwohl die Sauerstoff-Atome auch eine HBA-Eigenschaft aufweisen (vgl. Abbildung 10.2a: hier wird die Carboxylgruppe am Pyrrolidin-Ring zu dem Mo am Hf umgewandelt). Dies gilt analog z.B. für protonierte aliphatische Amine, die als PI codiert werden, obwohl das Stickstoff-Atom auch eine HBD-Eigenschaft aufweist (vgl. Abbildung 10.2a: hier wird das primäre Amin am Ende der Alkylkette zu dem Nb am Zn umgewandelt).

Stereochemie wird während der RG-Umwandlung nicht berücksichtigt, sodass Stereoisomere den gleichen RG ergeben. Auch Informationen über Stellungsisomerie am Ring geht im resultierenden RG verloren. Da Stereoisomere jedoch als gleiche RGs codiert werden, die sogleich auch die MCSs zwischen diesen Molekülen darstellen, erscheinen diese Moleküle am selben Netzwerk-Knoten im späteren Netzwerk. Wenn die Moleküle wieder an den entsprechenden MCS-Knoten abgebildet werden, sind die Ursachen für z.B. stereochemisch-bedingte sprunghafte SARs leicht identifizierbar.

a) Umwandlung von PI/NI-Eigenschaften

Im ersten Schritt werden positiv und negativ ionisierbare Eigenschaften umgewandelt. Diese Reihenfolge ist wichtig, damit beispielsweise Tetrazole als NI erkannt werden, aber auch andere Ringsysteme, in die PI- oder NI-Eigenschaften partiell eingebunden sind, komplett umgewandelt werden. Dazu wird immer geprüft, ob ein Atom mit PI- oder NI-Eigenschaft gleichzeitig auch Teil eines Ringsystems ist.

Die Umwandlung erfolgt hierbei immer nach dem folgenden Schema. Zunächst werden alle Atome des Moleküls, die in dasselbe Pseudoatom umgewandelt werden sollen, ermittelt. Durch Nachbarschaftssuchen in dem molekularen Graphen werden benachbarte Atome mit gleicher Eigenschaft zu einer Gruppe zusammengefasst. Dann werden die entsprechenden Nachbaratome zu dieser Atomgruppe ermittelt, die Atomgruppe gelöscht und ein neues Pseudoatom, das die Eigenschaft dieser Atomgruppe repräsentiert, erstellt. Dieses

Pseudoatom wird dann mit den vorher ermittelten Nachbaratomen wieder durch Einfachbindungen verknüpft. Eine Einfachbindung zwischen zwei Pseudoatomen im späteren RG zeigt an, dass die codierten Eigenschaften im Molekül verbunden waren.

b) Umwandlung von Ringen unter Berücksichtigung von Aromatizität und HBA/HBD

Im nächsten Schritt werden SSSR-basiert Ringe umgewandelt. Mittels Eigenschaftsmatrix wird neben der Aromatizität geprüft, ob Atome des SSSR weitere H-Brücken-Eigenschaften aufweisen. Gibt es ein Atom mit HBA- und ein anderes Atom mit HBD-Eigenschaft, so wird der Ring mit der Eigenschaft HBDA annotiert. Exozyklische Doppelbindungen werden separat codiert. So wird eine Carbonylgruppe als eigenständiger HBA codiert.

Unter Berücksichtigung der Anellierungsstellen im ursprünglichen Molekül werden Pseudoatome, die zwei anellierte Ringe repräsentieren, über eine Doppelbindung anstelle einer Einfachbindung miteinander verknüpft (vgl. Naphthyl-Ring in Abbildung 10.2b, der zu Sc=Sc umgewandelt wird).

c) HBA/HBD/HBDA-Eigenschafts-Erkennung

Im folgenden Schritt werden dann H-Brücken-Akzeptor und H-Brücken-Donor und kombinierte HBDA Eigenschaften in azyklischen Molekülteilen umgewandelt. Primäre oder sekundäre Amide werden hierbei nicht als HBDA, sondern als separate HBA (für den Carbonyl-Sauerstoff) und HBD (für das Wasserstoff-Atom am Amid-Stickstoff) umgewandelt (vgl. Abbildung 10.2b).

d) Terminale Gruppen und Linker

Nachdem alle vorgenannten mit entsprechenden Eigenschaften annotierten Atome zu entsprechenden Pseudoatomen umgewandelt worden sind, bleiben entsprechende Linker-Atome und Atome von terminalen Gruppen übrig. Diese werden nachfolgend wie oben beschrieben als Zn-Pseudoatome codiert.

e) Überprüfung der Konsistenz der RGs

Nachdem die RG-Umwandlung abgeschlossen ist, wird der finale RG zur Kontrolle auf Konsistenz geprüft. Hierbei werden RGs (z.B. mit nicht-umgewandelten Atomen oder die aus mehreren nicht-verknüpften Teilen bestehen), die aufgrund von chemischer Besonderheiten, die in der Implementierung ggf. noch nicht berücksichtigt sind, rausgefiltert und in eine Extra-Datei geschrieben. So ist es ständig möglich, die RG-Umwandlungs-Routine zu optimieren und um Sonderfälle zu ergänzen. Makrozyklische Moleküle, die ebenfalls nicht zu RGs umgewandelt werden, werden ebenfalls in eine separate Datei geschrieben.

10.3. Schritt 2: Erzeugung eines RG-MCSs-Pools

Nachdem alle Datensatzmoleküle als reduzierte Graphen codiert sind, ist der nächste wichtige Schritt die paarweise Bestimmung der maximal gemeinsamen Substruktur für alle Datensatzmoleküle bzw. deren RGs. Der komplette Python-Quellcode für die MCS-Bestimmung findet sich in Abschnitt 27.2 im Anhang.

1.) RG-Duplikate-Filter

Um die Berechnungen der MCSs zu beschleunigen, wird vorher mittels kanonischer SMILES[124] auf RG-Duplikate geprüft. Dies ist insbesondere für größere Datensätze von Bedeutung. In den meisten analysierten Datensätzen konnte so die Anzahl an paarweisen Vergleichen deutlich reduziert werden (etwa 40 bis 50%) aufgrund des Vorhandenseins einer hohen Anzahl an Serien analoger Moleküle.

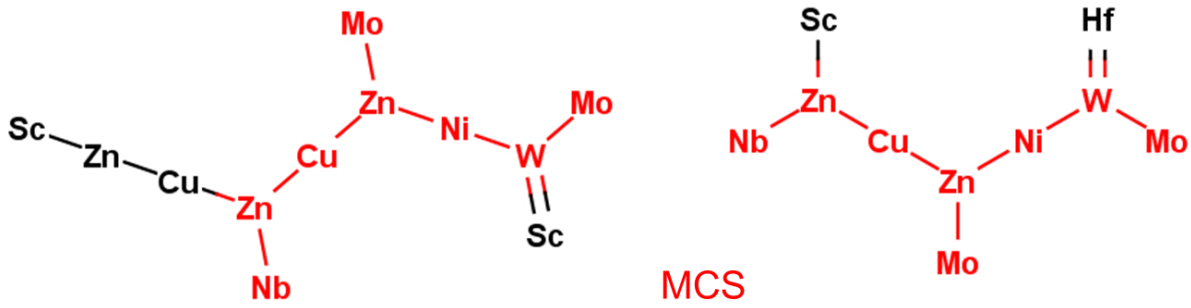


Abbildung 10.3. Beispiel für paarweise MCS-Bestimmung zwischen zwei RGs. Die maximal gemeinsame Substruktur ist rot hervorgehoben (MCS-Größe = 8).

2.) Paarweise MCS-Bestimmung

Im Folgenden werden dann alle paarweisen MCSs nur zwischen allen einzigartigen *z* RGs bestimmt (vgl. Abbildung 10.3) und in Form kanonischer SMILES in einer (*z* x *z*)-großen MCS-Matrix gespeichert. Gleichzeitig wird jeder neue MCS einer Liste einzigartiger MCSs hinzugefügt, sofern er dort noch nicht gespeichert ist.

Als MCS-Definition wird aus den in Kapitel 4.1 genannten Gründen der MCES anstelle des MCIS verwendet. Zudem wird nur der zusammenhängende MCS bestimmt, da sich hiermit einfacher eine gut interpretierbare, hierarchische Netzwerkstruktur erzeugen lässt als mit nicht-zusammenhängenden MCSs. Für die Bestimmung des MCS werden die MCS-Routinen des OEChem TK[370] verwendet, wobei in inSARa standardmäßig ein exakter Algorithmus zur Bestimmung des MCES verwendet wird. Optional kann bei sehr großen Datensätzen aber auch ein approximativer Algorithmus benutzt werden, um die Berechnungen zu beschleunigen. Da in der Regel nur bei komplexen und annelierten Ringsystemen größere Unterschiede auftreten, die aber bei inSARa durch einzelne

Pseudoatome zusammengefasst sind, sind keine großen Unterschiede zu erwarten. Obligat ist zudem eine MCS-Mindest-Größe von 3 Atomen. Optional kann diese zur Erhöhung der Spezifität der resultierenden MCSs erhöht werden (z.B. auf 5 Atome).

Wenn mehrere gleichgroße MCSs für ein RG-Paar (RG Nr. X und Y) gefunden werden (z.B. bei nicht-zusammenhängenden gemeinsamen Molekülteilen), wird keine Priorisierung dieser MCSs vorgenommen. Stattdessen werden alle größten gemeinsamen Substrukturen in der MCS-Matrix an der Position [X, Y] bzw. [Y, X] gespeichert bzw. in die Liste einzigartiger MCSs aufgenommen. Ein RG, der für mehrere Moleküle codiert, stellt ebenfalls einen MCS dar. Da diese Duplikate vorher gefiltert wurden, muss dieser MCS manuell der Gesamtmenge an MCSs hinzugefügt werden. Hierzu wird dieser Duplikate-RG auf der Diagonalen der MCS-Matrix an der zu dem RG gehörenden Position gespeichert bzw. in die Liste einzigartiger MCSs aufgenommen.

Die Substrukturen, die in dieser Matrix gespeichert sind, stellen für mindestens jeweils ein Molekülpaar MCSs dar. Für andere Moleküle handelt es sich bei diesen Strukturen jedoch nur eine gemeinsame Substruktur (Abk. CS, engl. common substructure), nicht aber die größte. Nichtdestotrotz werden diese CSs im Folgenden einfachheitshalber dennoch als MCSs bezeichnet.

10.4. Schritt 3: Aufbau der hierarchischen Netzwerk-Struktur

Das Kernstück der inSARa-Methode ist die Erzeugung einer hierarchischen Netzwerk-Struktur basierend auf der in Schritt 2 gewonnen Gesamtmenge an MCSs. Die komplette Netzwerk-Erzeugung wurde in Python implementiert unter Verwendung von Routinen des kommerziell erhältlichen OEChem TK[370] und der freiverfügbaren Programmierbibliotheken NetworkX[375–376]. Der komplette Quellcode findet sich in Abschnitt 27.3 im Anhang.

1.) Bestimmung potentieller Wurzel-MCSs

Der erste Schritt ist die Bestimmung von potentiellen Wurzel-MCSs aus der Gesamtmenge der MCSs, die wie unter 10.3 beschrieben in der Liste einzigartiger MCSs gespeichert sind. Ein „Wurzel-MCS“ ist definiert als MCS, der nur eine Substruktur, aber in keinem Fall eine Superstruktur irgendeines MCS in der Liste einzigartiger MCSs darstellt. Dieser Wurzel-MCSs wird im resultierenden baumartigen Netzwerk als „Wurzel-Knoten“ bezeichnet, weil er im Gegensatz zu normalen „MCS Knoten“ nur Nachfolger aber keine Vorgänger aufweist (vgl. Kapitel 3). Für die Erstellung der Liste potentieller Wurzel-MCSs wird für jeden MCS_A aus der Liste einzigartiger MCSs geprüft, ob kein MCS_B in dieser Liste existiert, der eine Substruktur des MCS_A darstellt. Wenn kein Substruktur-MCS gefunden wird, wird MCS_A der Liste potentieller Wurzel-MCSs hinzugefügt.

2.) Auswahl der Wurzel-MCSs

Als nächstes beginnt die Auswahl der Wurzel-MCSs aus der Menge potentieller Wurzel-MCSs. Hierzu wird aus der Liste potentieller Wurzel-MCSs immer derjenige MCS als Wurzel-MCS ausgewählt, der am meisten Moleküle des Datensatzes repräsentiert (d.h. dieser RG-MCS stellt eine Substruktur der RGs dieser Moleküle dar), die nicht bereits durch zuvor ausgewählte Wurzel-MCSs repräsentiert werden. Falls für mehrere MCSs die gleiche Anzahl an noch nicht repräsentierten Molekülen ermittelt wird, wird derjenige MCS bevorzugt und in der Liste der Wurzel-MCSs gespeichert, der weniger redundante Information (d.h. weniger schon bereits durch andere MCSs repräsentierte Moleküle) repräsentiert. Sollte durch diese Regel ein Gleichstand immer noch nicht aufgelöst werden können, so wird zufällig einer der in Frage kommenden MCSs ausgewählt. Dieser Auswahl-Prozess wird so lange wiederholt, bis ein benutzerdefinierter Anteil an Molekülen durch die ausgewählten Wurzel-Knoten im späteren Netzwerk repräsentiert ist (z.B. ≤ 2% nicht-repräsentierte Datensatz-Moleküle = Standard-Einstellung).

Falls der Datensatz strukturell zu heterogen ist oder die vorausgesetzte Mindest-MCS-Größe zu groß gewählt wird, kann es vorkommen, dass eine große Anzahl an Molekülpaaren keinen MCS der Mindestgröße aufweist. In diesen Fällen kann es passieren, dass kein weiterer Wurzel-MCSs in der Menge potentieller Wurzel-MCSs gefunden werden kann, der weitere noch nicht bereits repräsentierte Moleküle repräsentiert, obwohl das Stopp-Kriterium für das Beenden des Auswahlprozesses noch nicht erfüllt ist. In diesem Fall wird der Wurzel-Knoten-Auswahl-Prozess automatisch gestoppt. Des Weiteren kann der Auswahlprozess auch beendet werden, wenn eine benutzerdefinierte Anzahl an Wurzel-MCSs ausgewählt worden ist.

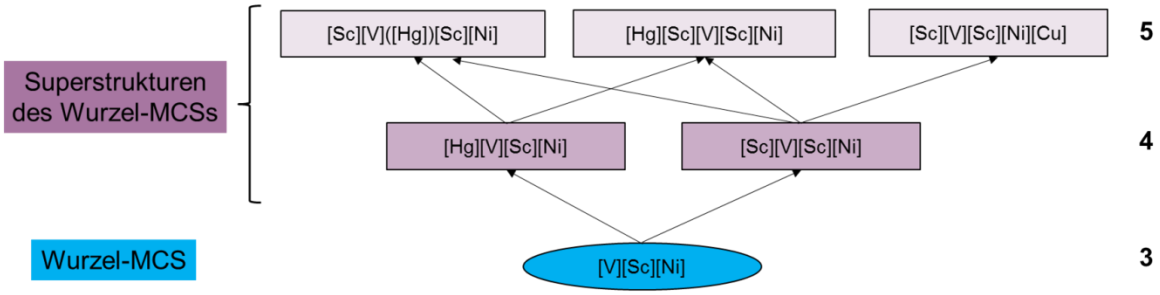


Abbildung 10.4. Schematische Darstellung der hierarchischen Netzwerk-Beziehungen zwischen einem Wurzel-MCS (blau) und denjenigen MCSs (lila), die Superstrukturen dieses Wurzel-MCS darstellen. Der Wurzel-MCS hat die Komplexitätsstufe 3 (Anzahl an Pseudoatomen), die Superstruktur-MCSs die Komplexität 4 und 5. Der MCS mit der geringeren Komplexitätsstufe ist immer mit denjenigen MCSs der nächsthöheren Stufe verbunden, von denen er eine Substruktur darstellt.

3.) Erzeugung einer hierarchischen Netzwerk-Struktur

Nach der Auswahl der Wurzel-MCSs wird durch iterative Sub- und Superstruktur-Suchen die hierarchische Netzwerkstruktur erstellt. Der Pseudocode für diesen Algorithmus ist in

Abbildung 10.5 dargestellt. Als Input für diesen Algorithmus wird neben der zuvor erzeugten Liste der Wurzel-MCSs eine modifizierte Liste einzigartiger MCSs benötigt. Diese wird durch das Entfernen aller potentieller Wurzel-MCSs von der ursprünglichen Liste der einzigartigen MCSs erhalten.

Der erste Schritt ist das Herausgreifen eines Wurzel-MCS aus der Liste finaler Wurzel-MCSs und das Erstellen eines Wurzel-Knoten in dem bisher leeren Netzwerk G. Dieser Wurzel-MCS repräsentiert die niedrigste Komplexitätsstufe bzw. Hierarchieebene in der Netzwerk-Hierarchie. Die Hierarchieebene wird durch die Anzahl an Pseudoatomen des zugehörigen MCS bestimmt. Die MCSs gleicher Komplexität in der modifizierten Liste einzigartiger MCSs werden zu Submengen zusammengefasst und nach steigender Komplexität sortiert.

Im nächsten Schritt werden durch Superstruktur-Suchen in der modifizierten Liste der einzigartigen MCSs alle MCSs identifiziert, die Superstrukturen dieses Wurzel-MCS darstellen und gleichzeitig die nächsthöhere Komplexitätsstufe (Größe des Wurzel-MCS plus ein Pseudoatom) repräsentieren. Für jeden dieser MCSs wird im Netzwerk G nachfolgend ein zugehöriger MCS-Knoten erstellt. Diese Knoten stellen zu diesem Zeitpunkt der Netzwerk-Erzeugung die terminalen Netzwerk-Knoten dar.

Für jeden MCS_k aus dieser MCS-Menge werden dann abermals durch Substruktur-Suchen im restlichen gesamten Netzwerk (Standard-Einstellung) oder nur in der aktuellen, zu dem Wurzel-Knoten gehörenden Komponente (optional) alle MCSs identifiziert, die eine Substruktur von dem MCS_k sind. Diejenigen MCSs bzw. deren zugehörige Knoten, die der höchsten Komplexitätsstufe angehören, werden anschließend mit dem MCS-Knoten des MCS_k im Netzwerk verbunden.

Im Folgenden wird die hierarchische Struktur aufgebaut durch iterative Erhöhung der in den Superstruktur-Suchen berücksichtigten Komplexitätsstufe (inkrementelle Erhöhung um jeweils ein Pseudoatom). Das bedeutet, dass aus der modifizierten Liste der einzigartigen MCSs im Verlauf der Implementierung MCSs mit zunehmender Größe in den Superstruktur-Suchen Berücksichtigung finden. Dieser Prozess wird solange wiederholt, bis alle Wurzel-MCSs aus der Liste finaler Wurzel-MCSs abgearbeitet sind. Am Ende ist jeder MCS bzw. MCS-Knoten mit den nächst-größeren MCSs, die eine Superstruktur von diesem kleineren MCS darstellen, verbunden. Dadurch wird eine in Abbildung 10.4 schematisch dargestellte, hierarchische Netzwerk-Struktur erhalten. Wie ebenfalls zu sehen ist, kann ein MCS-Knoten die Substruktur von mehreren MCS-Knoten sein. Dadurch wird die resultierende Netzwerk-Struktur sehr komplex und schwer interpretierbar.

Algorithm: Create hierarchical network structure

Input:

final root-MCS list = { root-MCS_1, root-MCS_2, root-MCS_3, …, root-MCS_n }

modified list of unique MCSs = { $subset_a$ = {all MCSs of size a}, $subset_{a+1}$, $subset_{a+2}$, …, $subset_m$}

option = False (= default → each MCS is represented once in the network, connections between different root-MCS allowed; option = True → each root-MCS and corresponding superstructure-MCSs as single disconnected subnetworks, multiple occurrence of MCSs allowed)

Output: hierarchical network structure G

```
G ← ∅
foreach i=1 to n do
    pick root MCS_i from final root-MCS list;
    a := size of root-MCS_i;
    create node N (G ← N);
    terminal node := N;
    foreach j=a+1 to m do
        foreach MCS_k in subset_j do:
            if root-MCS_i is substructure of MCS_k then
                create node M (G ← M);
                terminal node := M;
                mark all nodes as invalid;
                if option is False then
                    foreach existing node_l in current network with MCS_l of size < j do
                        if MCS_l is substructure of MCS_k then
                            mark the node_l as valid;
                if option is True then
                    foreach existing node_l in current component with MCS_l of size < j do
                        if MCS_l is substructure of MCS_k then
                            mark the node_l as valid;
                max-size := determine maximum MCS-size of all valid nodes;
                foreach node_o in the set of valid nodes do
                    if size of MCS of node_o < max-size then
                        mark node_o as invalid;
                create edge E (G ← E) from terminal node M to all valid nodes;
save network G;
```

Abbildung 10.5. Pseudocode (englisch) für den Algorithmus zur Erzeugung einer MCS-basierten hierarchischen Netzwerk-Struktur.

4.) Überführung des Netzwerkes in eine Baumstruktur: MST-Bestimmung

Um die Netzwerk-Struktur zu vereinfachen, wird daher der minimale Spannbaum (MST) des Netzwerkes berechnet (vgl. Kapitel 3). Hierfür wird der Kruskal-Algorithmus[223] (vgl. Kapitel 3) verwendet. Das Prinzip dieser Netzwerk-Prozessierung wird kurz in Abbildung 10.6 erläutert. Zwischenschritte, die von (b) nach (c) führen, werden in Abbildung 3.1 und Abbildung 3.2 in Kapitel 3 detailliert erläutert.

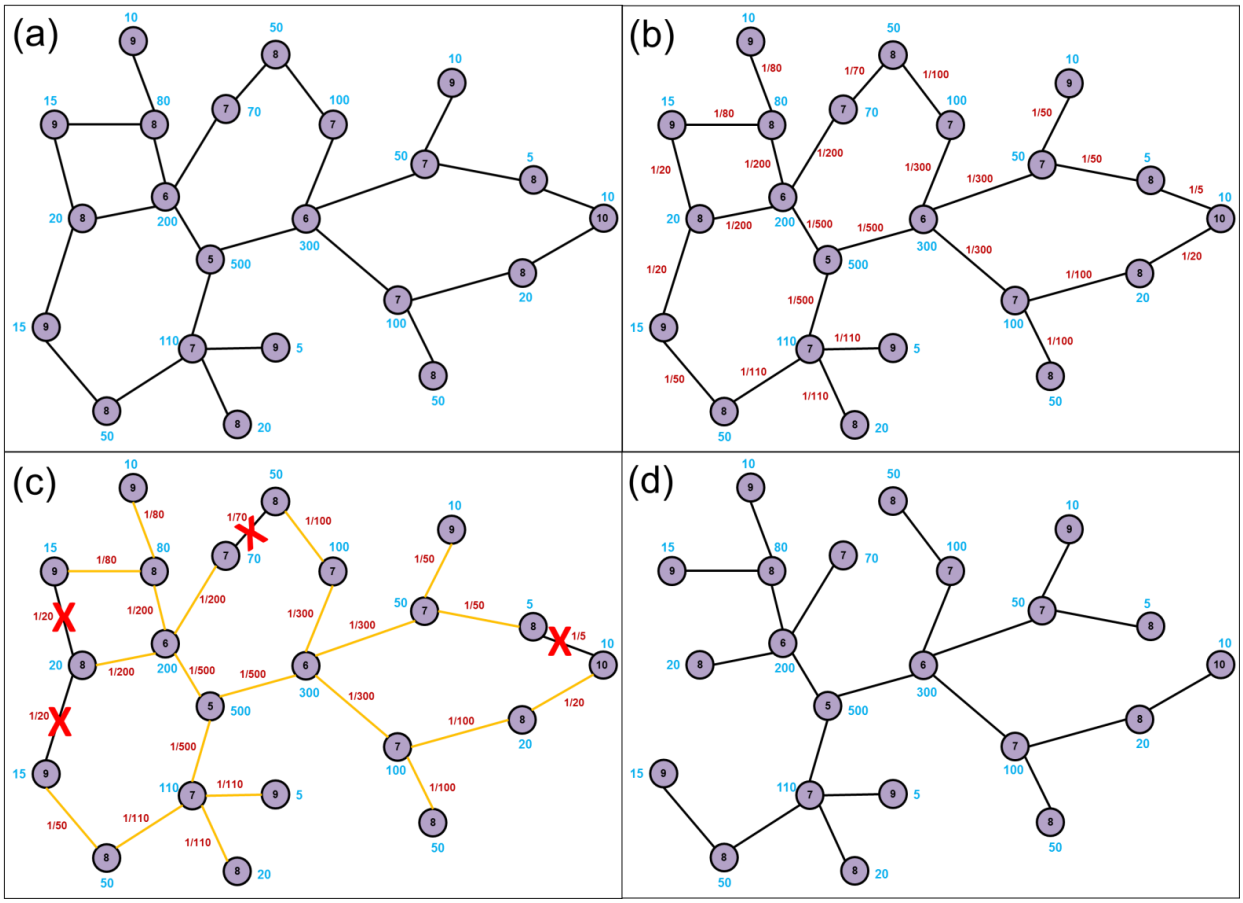


Abbildung 10.6. Prinzip der Reduktion der Netzwerk-Komplexität durch die Bestimmung des minimalen Spannbaumes (MST). **(a)** Schematische Darstellung des komplexen hierarchischen Ausgangsnetzwerkes. Die lila Knoten repräsentieren die MCS-Knoten. Die schwarze Zahl in den Knoten gibt die Komplexitätsstufe bzw. Hierarchieebene des zugehörigen MCS an. Die blaue Zahl neben dem jeweiligen Knoten zeigt die Anzahl an Molekülen an, die durch den MCS repräsentiert werden. **(b)** Zuordnung des Kantengewichtes (rot) zu den Kanten. **(c)** Bestimmung des MST (gelb markiert, für Details bezüglich der Zwischenschritte zwischen (b) und (c) vgl. Abbildung 3.1 und Abbildung 3.2 in Kapitel 3). Kanten, die nicht Bestandteil des MST sind, aber im ursprünglichen Netzwerk (a) vorkommen, sind mit roten Kreuzen gekennzeichnet. **(d)** Nach der MST-Bestimmung: Prozessiertes inSARa-Netzwerk mit baumartiger Struktur unter Erhalt der hierarchischen MCS-Beziehungen. Für weitere Details siehe Text.

In (a) wird das im vorgegangen Schritt erhaltene komplexe hierarchische Netzwerk gezeigt. In Schritt (b) wird die Zuordnung der Kantengewichte zu den Kanten zwischen den MCS-Knoten veranschaulicht. Als Kantengewicht wird die inverse Anzahl an repräsentierten Molekülen des inzidenten MCS-Knoten mit der geringeren Komplexitätsstufe verwendet.

Dieses Vorgehen stellt sicher, dass bei der MST-Erstellung die hierarchischen Beziehungen zwischen den MCSs erhalten bleiben. Andere Gewichte könnten unter bestimmten Umständen dazu führen, dass statt der in (c) mit roten Kreuzen angedeuteten Kanten andere Kanten bei der MST-Bestimmung geschnitten werden, um Kreise zu vermeiden. Bliebe z.B. die Kante mit dem Gewicht 1/70 erhalten und würde die benachbarte Kante mit dem Gewicht 1/200 wegfallen, würde auch ein kreisfreier Graph erhalten werden. Es würde jedoch ein terminaler Knoten der Komplexitätsstufe 7, der mit einem Knoten der Stufe 8 verknüpft ist, erhalten werden. Die hierarchische Struktur, bei der die Komplexität der MCSs von innen nach außen fortlaufend zunimmt, würde in diesem Fall verloren gehen. In (c) sind der resultierende MST (gelb) und nicht im MST vorkommende Kanten (rote Kreuze) dargestellt. Die MST-Bestimmung wandelt die komplexe Netzwerk-Struktur aus (a) in ein Netzwerk mit Baum-Struktur (d) um. Dies hat den Vorteil einer klaren hierarchischen Organisation und einfacherer Interpretierbarkeit.

5.) Abbildung der Moleküle im Netzwerk

Im letzten Schritt, um das endgültige inSARa-Netzwerk zu erhalten, werden die Datensatz-Moleküle an den zugehörigen MCS-Knoten abgebildet. Dazu wird jeweils geprüft, ob der MCS eine Substruktur des RGs des Moleküls darstellt. Bei der Visualisierung werden die Moleküle nur an den größten MCS-Knoten eines jeden Astes des Netzwerkes gezeigt, um die Netzwerk-Interpretation zu erleichtern. Moleküle können mehrfach im Netzwerk erscheinen. Nachteile dieses Vorgehens sind eine erhöhte Netzwerk-Komplexität und das Risiko redundanter Information. Jedoch kann das mehrfache Auftreten zum besseren Verständnis der chemischen Nachbarschaft beitragen und es ermöglicht ein bestimmtes Molekül in unterschiedlich struktureller Umgebung zu sehen.

Netzwerk-Varianten

inSARa-Netzwerke bestehen aus einer oder mehreren Baum-Strukturen, d.h. sie stellen einen Wald dar. Bei der Erstellung der hierarchischen Netzwerk-Struktur sind jedoch zwei Varianten möglich, die die Netzwerk-Komplexität und Informations-Redundanz stark beeinflussen. In der Standardeinstellung werden nicht-zusammenhängende Bäume über gemeinsame RG-MCSs verknüpft. Dieser Standard-Typ eines inSARa-Netzwerkes vermeidet Redundanz im Netzwerk und erleichtert das Erkennen von Gemeinsamkeiten zwischen verschiedenen Bäumen. Jedoch führt es auch zu größeren, komplexeren Bäumen. In der optionalen Variante führt jeder Wurzel-Knoten zu einem einzelnen Baum. Das hat den Nachteil, dass in verschiedenen Bäumen redundante Information gezeigt wird. Jedoch führt diese Variante normalerweise zu kleineren, weniger komplexen Sub-Netzwerken. Alle Beispiel-Netzwerke, die in Kapitel III gezeigt und diskutiert werden, wurden mit der Standard-Einstellung erstellt.

10.5. Schritt 4: Visualisierung der Netzwerke

Für die Visualisierung der inSARa-Netzwerke wird die Open-Source Netzwerk-Analyse- und Netzwerk-Visualisierungs-Software Cytoscape[377–380] genutzt. Zunächst werden die erzeugte Netzwerk-Datei (Format: ***.gml) und die zugehörigen Attribut-Dateien für MCS- und Molekül-Knoten (Format: ***.txt) importiert und anschließend ein Layout erzeugt. Aus Gründen der Übersichtlichkeit wird dafür der “force-directed layout” Algorithmus mit optimierten Parametern verwendet (siehe Tabelle 26.1 im Anhang). Bei kraftbasierten Layout-Algorithmen werden die Knoten als sich abstoßende Massen und die Kanten zwischen diesen wie Federn behandelt. Durch Simulation physikalischer Grundgesetze lassen sich übersichtliche Layouts erzeugen[381]. Bei sehr großen Netzwerken kann es trotz hoher Anzahl an Optimierungs-Iterationen zu Kantenkreuzungen kommen. Ein Vorteil von Cytoscape ist hierbei, dass das Netzwerk nach der Layout-Erstellung manuell nachbearbeitet werden kann. Kantenlängen oder Entfernungen zwischen nicht verbundenen Knoten sind auf den Layout-Algorithmus zurückzuführen. Wie in allen mathematischen Graphen haben sie keine chemische Bedeutung. Entscheidend ist allein die An- oder Abwesenheit von Kanten zwischen den einzelnen Knoten.

Durch die Verwendung des Cytoscape Plug-in chemViz[382–383] können zudem die 2D-Strukturen der Moleküle, sowie die reduzierten Graphen oder RG-MCSs direkt an dem zugehörigen Knoten abgebildet werden. Hierdurch wird die interaktive Interpretation der SARs zusätzlich erleichtert (vgl. Abbildung 10.7b).

10.6. Schematischer Aufbau von inSARa-Netzwerken

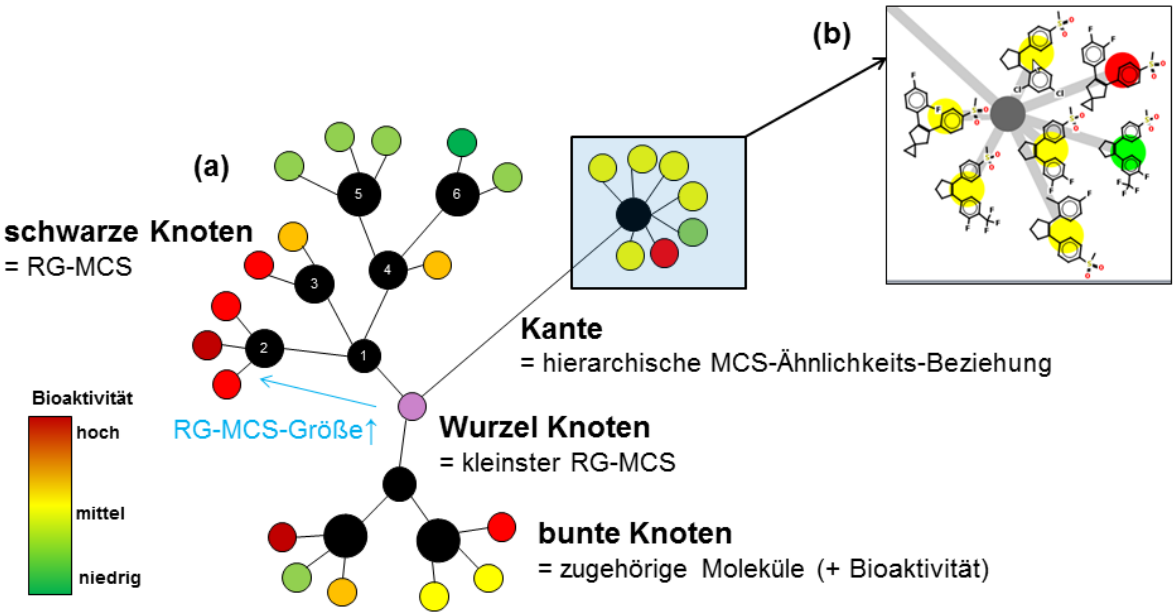


Abbildung 10.7: Prinzipieller Aufbau eines inSARa Netzwerkes. (a) Schematische Darstellung eines Muster-Netzwerkes, Erläuterungen siehe Text.(b) Mittels chemViz Plugin können die entsprechenden Datensatz-Moleküle nach dem Laden der erzeugten Netzwerk-Dateien in Cytoscape direkt an den zugehörigen Netzwerk-Knoten abgebildet werden.

In Abbildung 10.7 ist schematisch der Prototyp eines inSARa-Netzwerkes dargestellt. Die verschiedenen Knotentypen in den Netzwerken können anhand der Farbe unterschieden werden. Folgende Knotentypen treten im Netzwerk auf: Wurzel-Knoten, die jeweils einen Wurzel-MCS repräsentieren, MCS-Knoten zur Repräsentation von RG-MCSs und Molekül-Knoten, die jeweils ein Datensatz-Molekül darstellen. Wurzel- und MCS-Knoten sind schwarz gefärbt, wohingegen die Molekül-Knoten in Abhängigkeit von der Bioaktivität der zugehörigen Moleküle eingefärbt sind. Zur Vereinheitlichung wird dasselbe Farbspektrum (von Grün via Gelb nach Rot) verwendet, das bereits in anderen Programmen (z.B. SARANEA[136]) verwendet wurde. Grüne Knoten repräsentieren schwach aktive, gelbe Knoten mittelaktive und rote Knoten hoch aktive Moleküle. Die Schwellenwerte für die Bioaktivitäts-Grenzen können jeweils manuell an die zu analysierende Zielstruktur angepasst werden.

Um das Navigieren und Orientieren in den z.T. sehr komplexen Netzwerken darüberhinausgehend zu erleichtern, wird die Zunahme der RG-MCSs-Größe in den einzelnen Netzwerk-Ästen durch die Zunahme der Knotengröße der MCS-Knoten visualisiert. Kanten zwischen den einzelnen Knoten deuten an, dass zwischen den Knoten jeweils eine Beziehung besteht. So bedeuten Kanten zwischen Molekül-Knoten und einem MCS-Knoten, dass dieser RG-MCS eine Substruktur der jeweiligen RGs der entsprechenden Moleküle darstellt. Ebenfalls aus Gründen der Übersichtlichkeit werden die Moleküle immer nur am größten Substruktur-MCS-Knoten in dem jeweiligen Netzwerk-Ast abgebildet. Das bedeutet, dass die RGs der Moleküle jedoch nicht nur am abgebildeten MCS-Knoten eine Superstruktur des MCS darstellen, sondern auch an allen Vorgänger-MCS-Knoten.

Da inSARa-Netzwerke hierarchisch organisiert sind, zeigen Kanten zwischen MCS-Knoten eine Zu- oder Abnahme der Größe des RG-MCSs um ein oder mehr Pseudoatome an. Daher existieren die folgenden Beziehungen zwischen den MCS-Knoten des Netzwerkes in Abbildung 10.7: Der MCS des Wurzel-Knoten stellt eine Substruktur der MCSs aller weiteren MCS-Knoten in diesem Netzwerk dar. Das bedeutet, dass z.B. die MCS-Knoten 1, 2, 3, 4, 5 und 6 Superstrukturen des MCS des Wurzel-Knoten repräsentieren. Der MCS des Knoten 1 stellt wiederum eine Substruktur der von Knoten 2, 3 und 4 repräsentierten MCS dar und die MCSs von Knoten 5 und 6 stellen Superstrukturen des MCS von Knoten 4 dar. Von innen, also dem Wurzel-Knoten, nach außen, also den terminalen MCS-Knoten, nimmt im Netzwerk die MCS-Größe aufgrund des hierarchischen Aufbaus kontinuierlich zu. Dies hat zur Folge, dass die Moleküle, die mit einem bestimmten MCS-Knoten verbunden sind, umso ähnlicher sind, desto weiter außen sie sich im Netzwerk befinden bzw. desto größer der zugehörige MCS(-Knoten) ist. Die Ähnlichkeit zwischen den Molekülen ist an den äußersten MCS-Knoten also größer und SARs folglich leichter zu interpretieren.

11. Verwendete Datensätze

11.1. Kleinere QSAR-Datensätze

Für die Entwicklung von inSARa wurden QSAR-Datensätze von Sutherland et al.[384] und Fontaine et al.[385] verwendet. Diese haben den Vorteil, dass sie nur aus wenigen hundert Molekülen bestehen und somit die resultierenden Netzwerke entsprechend klein sind, sodass die Qualität der erzeugten inSARa-Netzwerke visuell schnell überprüft werden kann. Auch sind die Datensätze strukturell homogener zusammengesetzt und enthalten i.d.R. eine Reihe von Molekülen mit analoger Struktur.

11.2. Große Datensätze aus BindingDB

Da inSARa für die Anwendung auf Datensätze, wie sie z.B. in der Leitstruktur-Optimierung in der pharmazeutischen Industrie vorliegen, entwickelt wurde, müssen Datensätze größerer Dimensionen und heterogenerer Zusammensetzung für die Validierung der Methode verwendet werden. Um die Anwendbarkeit und Leistungsfähigkeit von inSARa für die Analyse von SARs in großen Datensätzen, die aus mehreren hundert bis tausend Molekülen bestehen, zu validieren, wurden daher verschiedene Datensätze, die jeweils aus mehr als 1000 Molekülen bestehen, aus der unter Abschnitt 2.2 beschriebenen, öffentlich freizugänglichen BindingDB zusammengestellt. Verglichen mit typischen unter 11.1 genannten QSAR-Datensätzen sind diese Datensätze nicht nur erheblich größer, sondern strukturell zumeist viel diverser. SAR-Interpretation ist somit deutlich anspruchsvoller.

Um die Beurteilung der Leistungsfähigkeit der inSARa-Methode bzw. der Qualität der resultierenden inSARa-Netzwerke zu vereinfachen, wurden für die Validierung Zielstrukturen mit bereits gründlich erforschten SAR-Charakteristika bevorzugt. Die sechs für tiefergehende Analysen ausgewählten Datensätze wurden ausgewählt, da die Zielstrukturen verschieden Target-Klassen angehören (fünf Enzyme (Kinasen, Proteasen, andere) und ein GPCR). Zudem weisen sie deutliche Unterschiede bezüglich Größe und struktureller Diversität auf. Eine Übersicht über die verwendeten Datensätze findet sich in Tabelle 11.1. Zusätzliche Datensatz-Charakteristika wie Bioaktivitäts-Verteilung, Fingerprint-Ähnlichkeiten und der globale SAR-Index[135] (berechnet mit SARANEA[136], vgl. Kapitel 2.4 und 2.6.4) finden sich in Tabelle 26.2 im Anhang.

Die Rohdaten für die Datensatz-Zusammenstellung wurden jeweils manuell als sdf-Datei (2D-Strukturen) Zielstruktur-spezifisch auf der Onlineplattform der BindingDB heruntergeladen. Diese Daten sind vorsortiert nach K_i- und IC_{50}-Bioaktivitätsdaten verfügbar. Eine nochmalige Filter-Prozedur hat sich jedoch als notwendig erwiesen.

Tabelle 11.1. Überblick über verwendete Datensätze aus der BindingDB.

Abkürzung	Target	Target-Klasse	Bioaktivitäts-Typ	Anzahl an Molekülen (nach Download)	Anzahl an Molekülen nach Vorbereitung und RG-Erzeugung	Anzahl einzigartiger RG
FXA	Faktor Xa	Enzym (Protease)	pIC_{50}	1887	1736	912
COX2	Cyclooxy-genase-2	Enzym	pIC_{50}	4357	2349	1083
CB1	Cannabinoid Rezeptor 1	GPCR	pK_i	2712	1957	890
CDK2	Cyclin-dependent Kinase 2	Enzym (Kinase)	pIC_{50}	2557	1575	979
P38	MAP Kinase p38 alpha	Enzym (Kinase)	pIC_{50}	2937	2446	1409
THR	Thrombin	Enzym (Protease)	pKi	3540	2852	1731

Kriterien für die Datensatzzusammenstellung

Aus den in Abschnitt 2.2.2 genannten Gründen wurden in dieser Arbeit nur Daten aus Bindungs-Assays (IC_{50}- oder K_i-Wert) verwendet. So sollte sichergestellt werden, dass die zu analysierenden Moleküle tatsächlich Liganden an dem zu untersuchenden Target sind und der biologische Effekt nicht auf andere intermolekularen Interaktionen zurückzuführen sind. Da in der BindingDB eine gute Zusammenstellung dieser Art von Daten (vgl. Abschnitt 2.2) zu finden ist und die Daten in bereits vorgefilterter und standardisierter Form mit zahlreichen Zusatzinformationen (z.B. direkter Verweis zur Quelle) zur Verfügung gestellt werden, wurde diese öffentlich zugängliche Datenbank als Datengrundlage verwendet.

Bei der Auswahl der Moleküle für die Datensatz-Zusammenstellung wurden einige Einschränkungen getroffen:

1.) Es wurden nur Moleküle, für die mit K_i- oder IC_{50}-Werten verfügbar sind, in der weiteren Vorbereitungsprozedur und späteren Analysen berücksichtigt. Da K_i- und IC_{50}-Bioaktivitätswerte nicht direkt miteinander vergleichbar sind, wurden die Moleküle dieser beiden Bioaktivitätstypen in separaten Datensätzen aufbereitet.

2.) Es wurden nur Moleküle mit einer bestimmten Mindest-Bioaktivität (für die SAR-Analysen einzelner Targets: K_i- oder IC_{50}-Werte kleiner als 100 µM) für die Datensatz-Zusammenstellung berücksichtigt.

3.) Handelte es sich bei den Bioaktivitätsdaten um Schwellenwerte (angezeigt durch “>” oder “<”), wurden die zugehörigen Moleküle verworfen.

12. Datenvorbereitung

Um eine einheitliche molekulare Repräsentation aller Moleküle zu gewährleisten, wurden alle zu analysierenden Datensätze nach einem Standardprotokoll vorbereitet. Dies ist v.a. bei Datensätzen wichtig, die aus verschiedenen Datenquellen stammen. Die aus der BindingDB zusammengestellten Datensätze gehören wie unter 2.2.2 beschrieben auch zu dieser Form von heterogenen Daten, wo diese Art der Datenvorbereitung empfehlenswert ist. Zur Standardisierung wurden MOE's sdwash und sdfilter Funktionen[188], sowie selbstgeschriebene Python-Skripte verwendet.

Ausschluss-Filter mittels „sdfilter"

Mittels sdfilter wurden zuerst alle Moleküle mit einer Ringgröße größer oder gleich neun Ringatomen (Option „-smallring"), sowie alle Moleküle mit einer molaren Masse größer als 800 Dalton ausgeschlossen (Option „-mw 800-"). Dieser Masse-Filter wurde in Anlehnung an STIEFL et al.[239] und CHEN und REYNOLDS[386] gewählt. Ein solcher Schwellenwert ist sinnvoll, da sehr große Moleküle weniger „drug-like" sind[387–388] und Voranalysen zudem gezeigt haben, dass bei ihrer Codierung sehr große RGs resultieren. Da Analysen der Verteilung der molaren Masse bei der Datensatzvorbereitung ergeben haben, dass beispielsweise Moleküle aus der Gruppe der Protease-Inhibitoren (z.B. Thrombin- oder Faktor-Xa-Inhibitoren) oftmals ein höheres Molekulargewicht als die Standardgrenze von 500 oder 600 Dalton aufzuweisen, wurde 800 Dalton als Schwellenwert gewählt. Des Weiteren wurden nur Moleküle im Datensatz behalten, die aus den häufigen Elementen organischer Verbindungen bestehen: „-elements C, H, N, O, S, P, F, Cl, Br, I". Metalloorganische Verbindungen wurden nicht analysiert.

Molekül-Standardisierung mittels „sdwash"

Mittels sdwash wurden dann alle Gegenionen, Lösungsmittelmoleküle und andere mögliche Addukte entfernt, indem nur das größte Molekül-Fragment behalten wurde (Option „-component"). Zudem wurde der Protonierungszustand aller Moleküle an den physiologischen pH-Wert von 7,4 durch die regelbasierte Protonierung bzw. Deprotonierung von starken Basen bzw. Säuren (Option „-acidbase") angepasst. Dies ist wichtig, damit später bei der Umwandlung der Moleküle in reduzierte Graphen die ionischen pharmakophoren Eigenschaften richtig erkannt werden.

Filtern von Duplikaten

Nach dieser Vorbereitungsprozedur wurden im letzten Schritt Duplikate mittels isomerischer kanonischer SMILES (Implementierung des OEChem TK[370]) gefiltert. Existierten für ein Molekül mehrere Bioaktivitäts-Daten für dieselbe Zielstruktur, so wurde überprüft, wie stark diese Werte differieren. Analog zu WASSERMANN und BAJORATH wurde folgendermaßen verfahren[389]: Sofern die Werte um nicht mehr als eine log-Einheit voneinander abweichen, wurde der arithmetische Mittelwert aller Werte berechnet und als Grundlage für die späteren SAR-Analysen verwendet. Ansonsten wurde das entsprechende Molekül verworfen, um einen Fehler der Analysen, der auf inkonsistente Bioaktivitätsdaten z.B. aufgrund von Messfehler bei der Durchführung des Assays oder Fehlern bei der Extraktion der Werte aus der Primärquellen zurückgeht (vgl. Kapitel 2.2.2) zu vermeiden.

13. Netzwerk-Optimierung und Ähnlichkeits-Analyse

13.1. Analyse zur Identifizierung von unspezifischer RG-Ähnlichkeit

Um die Spezifität der in den Netzwerken enthaltenen Information (gemeinsame RG-MCSs) zu optimieren, wurde die Zufalls-Ähnlichkeit zwischen Molekülen analysiert. Ziel der Untersuchung war es zu ermitteln, welche RG-MCSs bei der MCS-Bestimmung zwischen zufällig gewählten Molekülpaaren resultieren. Im Gegensatz zu sonstigen analysierten Datensätzen, die aus biologisch aktiven Molekülen an einer Zielstruktur bestehen, stehen die Moleküle in keiner bekannten Beziehung zueinander. Es ist jedoch nicht auszuschließen, dass zufällig zwei Moleküle, die an demselben Target Aktivität zeigen, gezogen werden. Die Wahrscheinlichkeit ist jedoch aufgrund der Größe und Diversität der ausgewählten Datenbank gering. Die Ergebnisse aus dieser Analyse sollen dazu genutzt werden, unspezifische Information bei der Netzwerk-Erzeugung zu erkennen bzw. zu eliminieren, um die inSARa-Netzwerke spezifischer zu gestalten. Ein weiteres Ziel dieser Analyse war zu ermitteln, welche Mindest-Größe ein Wurzel-MCS haben sollte, um eine vernünftige Spezifität aufzuweisen.

Datengrundlage und Vorbereitung

Als Datengrundlage für die Analyse diente die ZINC Datenbank (Version 12)[390–392]. Hierbei handelt es sich um eine frei zugängliche Datenbank von über 21 Millionen kommerziell erhältlichen Molekülen, zusammengestellt aus über 150 verschiedenen Händler-Katalogen. Die Moleküle liegen annotiert und kategorisiert nach bestimmten Eigenschaften bzw. kommerzieller Verfügbarkeit/Lieferzeit in standardisierter Form in der Datenbank vor. Sie sind bereits zur Nutzung für das Virtuelle Screening (z.B. mittels molekularem Docking) aufbereitet. Für diese Analyse wurde nur eine Untermenge („Clean Drug-Like“) der gesamten ZINC Datenbank bestehend aus etwa 12 Millionen vorgefilterten Arzneistoff-ähnlichen Molekülen (Definition nach LIPINSKI[387]) verwendet. Diese Untermenge wurde ausgewählt, da die zugehörigen Moleküle repräsentativ, für die mit inSARa analysierten Datensätze aus der BindingDB sind. “Clean” bedeutet, dass Moleküle mit bestimmten problematischen funktionellen Gruppen (z.B. mit toxischen oder reaktiven Eigenschaften) verworfen wurden.

Die gesamten ZINC Moleküle wurden zunächst nochmals mittels MOE’s sdwash and sdfilter Funktionen[188] (vgl. Datenvorbereitung in Kapitel 12) standardisiert. Anschließend wurden Duplikate mittels InCHIs[125–126] (vgl. Abschnitt 2.3.2), die mit Open Babel[114–115] erzeugt wurden, unter Berücksichtigung von tautomeren Formen gefiltert. Diese vorbereiteten Moleküle wurden dann wie unter 10.2 beschrieben als reduzierte Graphen codiert. Nach der Filterprozedur und der Umwandlung in RGs, bei der einzelne Moleküle (z.B. mit zu großen Ringsystemen) ebenfalls wegfallen, beinhaltete die Datenbank etwa 11 Millionen Molekülen.

Versuchsaufbau und Auswertung

Zur Analyse der Zufalls-Ähnlichkeit wurden zufällig jeweils zwei Moleküle aus der zuvor beschriebenen, aufbereiteten Datenbank gezogen und anschließend der RG-MCS dieses Zufallspaares bestimmt. Die minimale MCS-Größe wurde auf 3 RG-Pseudoatome festgelegt. MCSs mit weniger als 3 Atomen wurden als kein gemeinsamer MCS gewertet. Um eine repräsentative Statistik über Zufalls-Ähnlichkeiten zu bekommen, wurde dieses Vorgehen eine Millionen Mal wiederholt.

13.2. Analyse der Korrelation zwischen FP- und MCS-basierter Ähnlichkeit

Zusätzlich zu dem RG-MCS wurden bei der unter 13.1 beschriebenen Analyse auch die Tc-Ähnlichkeiten für die beiden am häufigsten verwendeten Fingerprint-Typen ECFP4 und MACCS Keys (166 Bits) jeweils für ein Molekülpaar bestimmt, um zu untersuchen, ob es ein Zusammenhang zwischen MCS-Größe und Fingerprint-Ähnlichkeit existiert.

Zum Vergleich wurde für verschiedene Datensätze (FXa, CB1, COX2, P38, vgl. Tabelle 11.1) ebenfalls die RG-MCS-Größe und die zugehörige Tc-Ähnlichkeit für ECFP4 und MACCS Keys für alle Molekülpaare bestimmt.

Außerdem wurde die Korrelation zwischen verschiedenen Fingerprints und der RG-MCS-Größe bzw. der MCS-Ähnlichkeit (berechnet als RASCAL-Score, vgl. Abschnitt 2.3.3) berechnet. Als Fingerprints wurden ausgewählt: ECFP4, MACCS Keys als binäre (MACCS) und Integer-Variante (MACCSF), CATS2D und FP2 (für Details zu den Fingerprints vgl. Abschnitt 15.2). Hierfür wurde der Spearman-Rang-Korrelationskoeffizient ρ_S[393] verwendet, da bei diesem im Gegensatz zum häufig verwendeten Pearson-Bravais-Produkt-Moment-Korrelationskoeffizienten r_P[394] weder normalverteilte Daten noch ein linearer Zusammenhang vorausgesetzt werden muss[395]. Da diese Voraussetzungen bei den Daten nicht angenommen werden können, ist der ρ_S ein robusteres Maß für die Bestimmung der Korrelation. Zur Berechnung des ρ_S zwischen zwei Methoden x und y müssen zunächst alle paarweisen Ähnlichkeitswerte, die mit einer Methode bestimmt werden, in Ränge umgewandelt werden. Der für das Molekülpaar i mit Methode x bestimmte numerische Wert X_i und mit Methode y bestimmte numerische Wert Y_i wird dann jeweils basierend auf dem entsprechenden Rang in der sortierten Liste zugeordnet, d.h.: x_i = Rang(X_i) bzw. y_i = Rang(Y_i). Anschließend kann unter Verwendung dieser Ränge analog zum Pearson-Korrelationskoeffizienten der Spearman-Korrelationskoeffizient ρ_S wie folgt berechnet werden, wobei $\bar{x}$ bzw. $\bar{y}$ jeweils die Mittelwerte aller Ränge von x bzw. y darstellen:

$$\rho_S = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \qquad \textbf{(13.1)}$$

ρ_S kann Werte von -1 bis 1 annehmen, wobei das Vorzeichen die Richtung der Korrelation angibt. Ein Wert von 1 zeigt perfekte positive Korrelation zwischen den Rängen zweier

Methoden an, während -1 perfekte negative Korrelation angibt. Ein Wert von 0 zeigt an, dass keine Korrelation besteht.

13.3. Analyse weiterer Optimierungs-Parameter

Bei der Netzwerk-Erstellung können einige Parameter wie die Mindest-MCS-Größe und das Stopp-Kriterium bei der Wurzel-Knoten-Auswahl zum Fine-Tuning der Netzwerke optional variiert werden. Der Einfluss dieser Parameter auf die Netzwerk-Komplexität und -Topologie wurde anhand von Datensätzen unterschiedlicher Größe und struktureller Variabilität analysiert.

Datengrundlage

Für die Analyse wurden die sechs Datensätze aus Tabelle 11.1 verwendet. Nach der Vorbereitung und der Umwandlung in RGs bestanden diese Datensätze aus etwa 1500 bis 3000 einzigartigen Molekülen.

Versuchsaufbau und Auswertung

Für jeden dieser Datensätze wurden inSARa-Netzwerke wie in Abschnitt 10 beschrieben mit Standardeinstellungen erzeugt. In einer Analyse wurde die Mindest-MCSs-Größe von 3 bis 8 Pseudoatome, in einer weiteren Analyse das Abbruchkriterium von 1 bis 25% (Anteil nicht-repräsentierter Moleküle) variiert. Die resultierende Anzahl an Wurzel- und MCS-Knoten im Netzwerk, sowie der Anteil nicht-repräsentierter Moleküle und die Anzahl an resultierenden Komponenten wurden als Maßzahlen für die Netzwerk-Komplexität und -Topologie verwendet.

14. inSARa Hybrid: Kombination mit Fingerprints

14.1. Zielsetzung

Wie in Kapitel 4.4 beschrieben, kann eine Kombination des MCS-Konzeptes mit Fingerprints oftmals vorteilhaft sein (z.B. Reduktion des Rechenaufwandes der MCS-Bestimmung durch das Vor-Clustern mit FPs). Davon inspiriert ist der nachfolgend beschriebene inSARa Hybrid Ansatz entstanden, mit dem untersucht werden sollte, welche Vorteile aus inSARa- und Fingerprint-basierter SAR-Analyse zu erwarten sind. inSARa Hybrid stellt eine Kombination aus inSARa und Fingerprint-basierten Ähnlichkeits-Netzwerken dar. Es wurden zwei verschiedene Varianten der Kombination mit unterschiedlicher Zielsetzung entwickelt.

Variante A: Vereinfachung der FP-basierten Netzwerk-Analyse mittels inSARa bzw. Verwendung von inSARa zur Analyse großer Cluster

Die *Variante A* versucht verschiedene Probleme beider Netzwerk-Typen zu lösen. So haben Fingerprint-basierte Ähnlichkeits-Netzwerke den Nachteil, dass ihnen eine klare Struktur fehlt und schwer zu erkennen ist, wodrauf die Ähnlichkeit zwischen den einzelnen Molekülen beruht. inSARa-Netzwerke hingegen können aufgrund der RGs und des damit verbundenen Abstraktionslevels unter Umständen sehr abstrakte Beziehungen zwischen Molekülen herstellen. Zudem werden inSARa-Netzwerke mit steigender Datengröße unter Umständen sehr komplex (vgl. Analyse aus Kapitel 18.3).

Diese Variante kann als eine Spezial-Form des Datensatz-Clusterns angesehen werden, wobei inSARa zur anschließenden Analyse großer Cluster verwendet wird. Das Clustern der Moleküle erfolgt in diesem Fall durch das Erstellen eines Fingerprint-basierten Ähnlichkeits-Netzwerkes (vgl. NSGs in Kapitel 2.6.4). Als Cluster werden alle Moleküle einer Zusammenhangskomponente (vgl. Kapitel 3) definiert. Für Cluster einer definierten Mindest-Größe wird anschließend ein inSARa-Netzwerk erstellt. Dieser Hybrid-Ansatz bietet den Vorteil, dass zum einen die Netzwerk-Komplexität der resultierenden inSARa-Netzwerke reduziert werden kann, was die Interpretierbarkeit deutlich vereinfacht. Zudem wird das Risiko von zu abstrakten Gruppierungen im inSARa-Netzwerk durch Ausschluss zu unähnlicher Molekülpaare reduziert. Der Vorteil der Verwendung von Komponenten besteht darin, dass jedoch nicht nur Moleküle oberhalb eines bestimmten Ähnlichkeitsschwellenwert bezogen auf ein bestimmtes Molekül, sondern jeweils auch die nächsten Nachbarn dieser ähnlichen Nachbarmoleküle weiterhin für die SAR-Analyse zur Verfügung stehen. Dadurch können im inSARa-Netzwerk immer noch Beziehungen (gleiche pharmakophore Eigenschaften, unterschiedliche Molekülstruktur) zwischen bestimmten Molekülen hergestellt werden, die in NSG-ähnlichen Netzwerken nicht direkt deutlich werden. Die SAR-Analyse der ursprünglichen FP-basierte Netzwerkkomponente wird durch die hierarchische, auf klaren Substruktur-Beziehungen basierende inSARa-Netzwerk-Struktur deutlich vereinfacht.

Bei dem verwendeten Cluster-Verfahren handelt es sich um ein graphentheoretisches Cluster-Verfahren (vgl. KUMAR[396]). Die Cluster-Definition über die Zusammenhangskomponte ist im Vergleich zu anderen graphentheoretischen Cluster-Definitionen (wie z.B. über maximal verknüpfte Subgraphen/„Cliques“, vgl. Kapitel 3) in

Anbetracht der chemischen Bedeutung der zugrundeliegenden Graphenstruktur am vorteilhaftesten. Bei diesem Verfahren ist zu beachten, dass die Clusterbildung immer von dem Schwellenwert abhängig ist, der für die Bildung einer Kante zwischen zwei Knoten verwendet wird.[396] Das für inSARa Hybrid verwendete Cluster-Verfahren kann ebenfalls als eine Form des nächsten-Nachbarn-Clustering angesehen werden.[397] Andere Formen des nicht-hierarchischen Clusterings wie der häufig verwendete *k*-Means-Algorithmus[398] oder die künstlichen neuronalen Netze (engl. Artificial Neuronal Network, Abk. ANN)[399] bzw. selbstorganisierenden Karten/Kohonenkarten (engl. Self-Organising Map, Abk. SOM)[400] oder des hierarchischen Clusterings (vgl. Übersichtsartikel von JAIN et al.[397] für einen umfassenden Überblick) könnten ebenfalls verwendet werden. Ein Nachteil des *k*-Means besteht darin, dass vor der Anwendung des Algorithmus die Anzahl an Gruppen (*k*) (ggf. unter Verwendung bestimmter Heuristiken[401]) definiert werden muss.[398] Beim hierarchischen Gruppieren hingegen ist es notwendig am Ende des Verfahren einen Grenzwert zu definieren, der das erhaltene Dendrogram auf einer bestimmten Höhe abschneidet, um die Cluster zu erhalten.[397]

Variante B: Einschränkung der MCS-Bestimmung zur Reduktion der inSARa-Netzwerk-Komplexität

In Kapitel 18.3 werden einige Möglichkeiten zur Reduktion der Komplexität von inSARa-Netzwerken analysiert. Ein Faktor, der die Komplexität entscheidend mitbestimmt, ist die Gesamtmenge an MCSs (vgl. Abschnitt 10.3). Durch Einschränkungen für die MCS-Bestimmung kann diese Menge u.U. um kleinere oder weniger bedeutende MCSs reduziert werden. Bei dieser Variante B wird daher untersucht, inwieweit sich die Gesamtmenge an MCSs durch Berücksichtigung von FP-Ähnlichkeit bei der MCS-Bestimmung reduzieren lässt und welchen Einfluss dies auf die nachfolgende Erzeugung des inSARa-Netzwerkes hat.

14.2. Methode

In Abbildung 14.1 wird das Prinzip der verwendeten Fingerprint-basierten Ähnlichkeits-Netzwerke und des inSARa Hybrid Ansatzes (Variante A) zusammengefasst.

Schritt 1: Erzeugung von Fingerprint-basierten Ähnlichkeits-Netzwerken

Voraussetzung für beide Varianten des inSARa Hybrid Ansatzes ist die Erzeugung eines Fingerprint-basierten Ähnlichkeits-Netzwerkes. Dies wird nach dem Prinzip der Network-like Similarity Graphs[193] erstellt (vgl. Abschnitt 2.6.4). Hierfür wird jedes Datensatz-Molekül als Knoten dargestellt und eine Kante zwischen zwei Knoten gebildet, wenn ein vorher definierter Ähnlichkeits-Schwellenwert überschritten wird. Dieser Schwellenwert wird in Abhängigkeit vom verwendeten Fingerprint definiert (z.B. MACCS Keys: Tc-Ähnlichkeit 0.65-0.75[193–194]; ECFP4: 0.4-0.55[195]). Je höher der Schwellenwert gewählt wird, desto strukturell ähnlicher müssen zwei Moleküle sein, damit eine Kante im Netzwerk gebildet wird. Durch die Erhöhung des Tc-Schwellenwertes, erhöht sich daher auch die Zahl der entstehenden Zusammenhangskomponenten. Aufgrund der besten Korrelation mit MCS-

basierter Ähnlichkeit (vgl. Abschnitt 18.2) wurden die FP-basierten Ähnlichkeits-Netzwerke unter Verwendung des ECFP4-Fingerprints (MOE-Implementierung[402]) erstellt. Als Schwellenwert wurde ein Tc von 0.55 gewählt.

Alle in dieser Arbeit (in Teil III) gezeigten Fingerprint-basierten Ähnlichkeits-Netzwerke wurden mit selbst-geschriebenen Python-Skripten erzeugt. Für die Visualisierung wurde (analog zu den inSARa-Netzwerken) Cytoscape verwendet (vgl. Kapitel 10.5.). Als Layout wurde ebenfalls das „force-directed Layout“ verwendet (vgl. Kapitel 10.5.). Die Knoten wurden entsprechend der Bioaktiviät des zugehörigen Moleküls eingefärbt (Farbschema analog zu den NSGs und den Molekül-Knoten in den inSARa-Netzwerken, vgl. Kapitel 10.6).

Schritt 2:

Variante A: inSARa-Netzwerk-basierte Cluster-Analyse

Anschließend werden im FP-basierten Netzwerk die Zusammenhangskomponenten bestimmt. Alle Moleküle einer Zusammenhangskomponente werden anschließend in RGs umgewandelt und der paarweise RG-MCS zwischen allen Molekülen einer Zusammenhangskomponente bestimmt. Ausgehend davon wird dann ein hierarchisches inSARa-Netzwerk erstellt, das wie in Kapitel 16 beschrieben zur SAR-Interpretation verwendet werden kann.

Variante B: Reduktion der Gesamtmenge an MCSs

Das Fingerprint-basierte Ähnlichkeits-Netzwerk kann ebenfalls dazu verwendet werden, die Gesamt-Menge an MCSs, die die Grundlage bei der inSARa-Netzwerk-Erzeugung darstellt, zu reduzieren. Hier wurden zwei Varianten untersucht:

I.) **Komponenten-Variante:** Es wird nur noch der MCS zwischen Molekülen, die sich in der gleichen Zusammenhangskomponente des FP-basierten Ähnlichkeits-Netzwerkes befinden, bestimmt.

II.) **Adjazenz-Variante:** Um die MCS-Menge weiter einzuschränken, kann alternativ der MCS nur zwischen im FP-basierten Ähnlichkeits-Netzwerk adjazenten Molekülen bestimmt werden. Hierbei handelt es sich um Molekülpaare, bei denen die FP-Ähnlichkeit den vordefinierten Tc-Schwellenwert überschreitet und die daher durch eine Kante im Netzwerk miteinanderverbunden sind.

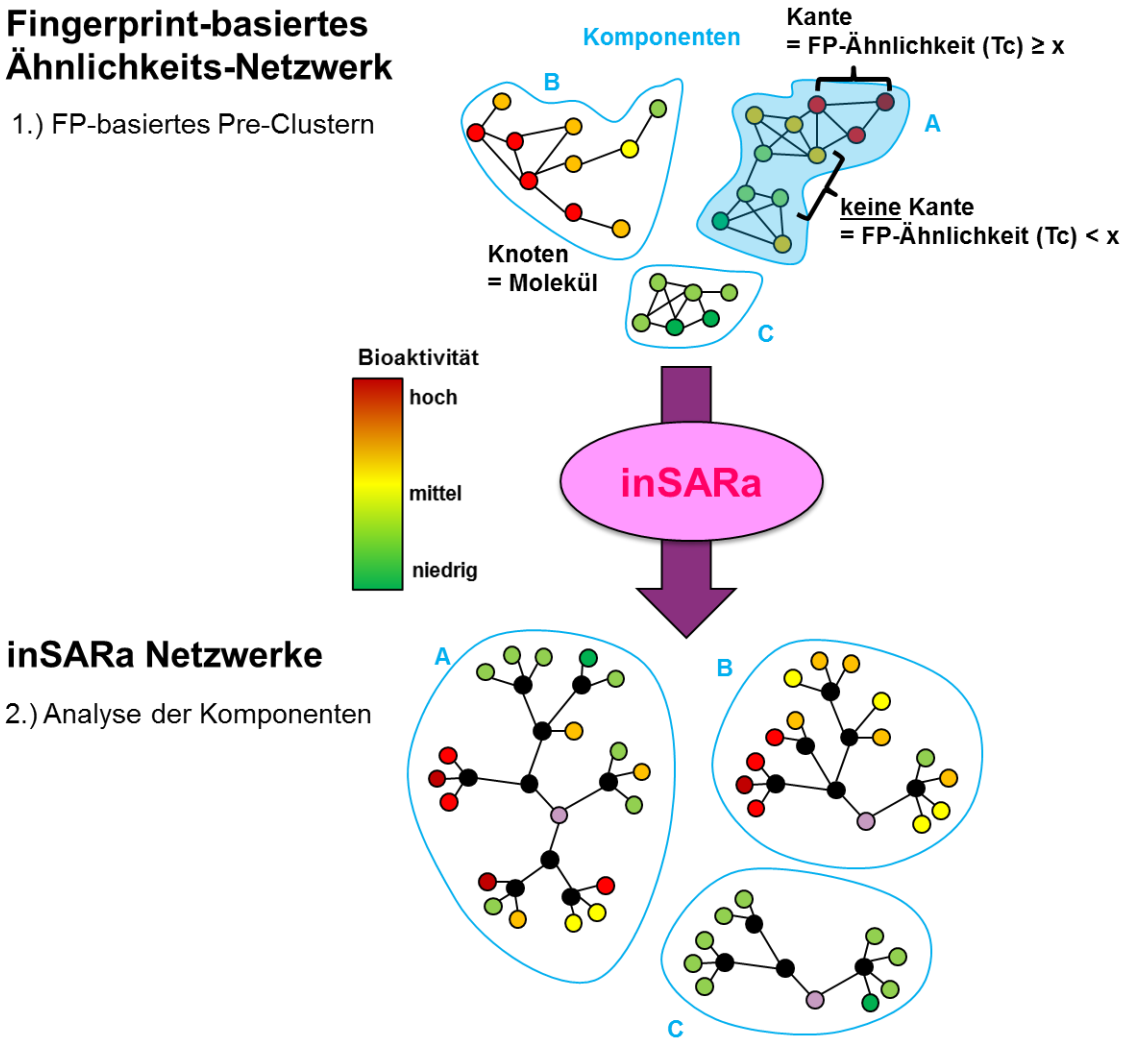


Abbildung 14.1. Prinzip der inSARa Hybrid Variante A. Im ersten Schritt wird ein Fingerprint-basiertes Ähnlichkeits-Netzwerk (vgl. NSGs) erzeugt. Dazu muss ein vom verwendeten FP-abhängiger Ähnlichkeits-Schwellenwert Tc definiert werden. Dieser bestimmt, ob zwischen zwei Molekül-Knoten eine Kante im Netzwerk erzeugt wird. Je nach struktureller Diversität entstehen mehrere Zusammenhangskomponenten. Zur Analyse der SARs in den einzelnen Zusammenhangskomponenten werden dann im zweiten Schritt hierarchische RG-MCS-basierte inSARa-Netzwerke erstellt.

14.3. Anwendung und Auswertung

Zur Anwendung wurden die größten drei Datensätze aus Kapitel 11.2 (THR, CB1 und P38) analysiert. Es wurden jeweils Fingerprint-basierte Netzwerke mit verschiedenen Tc-Schwellenwerten (0, 0.4, 0.45, 0.5, 0.55, 0.6, 0.7) erstellt. Auf Basis dieser Information wurde jeweils eine Komponenten-Matrix und eine Adjazenz-Matrix erzeugt, die für jedes Molekülpaar angibt, ob sich die beiden Moleküle in der gleichen Komponente im FP-basierten Netzwerk befinden bzw. adjazent sind (Wert „1“). Diese Matrizen wurden jeweils dazu verwendet, die ursprüngliche MCS-Matrix zu modifizieren. Ein MCS eines Molekülpaares XY bleibt nur in der MCS-Matrix, wenn in der Komponenten- oder Adjazenzmatrix der Wert für dieses Paar auf „1“ gesetzt ist. Anschließend wurde basierend auf dieser modifizierten MCS-Matrix jeweils ein inSARa-Netzwerk mit den folgenden Standardeinstellungen erstellt: Mindest-MCS-Größe = 5 RG-Atome, Abbruch-Kriterium = 2% nicht-repräsentierte Moleküle, Ausschlussliste = aktiv.

Zur Charakterisierung des Einfluss auf die Gesamt-MCS-Menge wird die Anzahl an einzigartigen MCSs in der modifizierten MCS-Matrix bestimmt (für Mindest-MCS-Größe: 3 und 5 RG-Atome). Zur Analyse des Einflusses auf das resultierende inSARa-Netzwerk wurden Maßzahlen zur Charakterisierung der Netzwerk-Komplexität und -Topologie, sowie der Bioaktivitätsverteilung an den MCS-Knoten („Knoten-Reinheit“) berechnet. Analog zu der in Abschnitt 13.3 beschriebenen Analyse wurde die resultierende Anzahl an Wurzel- und MCS-Knoten im Netzwerk, sowie der Anteil nicht-repräsentierter Moleküle und die Anzahl an resultierenden Komponenten als Maßzahl zur Charakterisierung der Komplexität und Topologie verwendet. Zur Charakterisierung der Knoten-Reinheit wurde für alle MCS-Knoten mit einer MCS-Ähnlichkeit ≥ 0.6 (vgl. MCS-Sim-Score in Kapitel 16.2) das für Ausreißer robuste Verteilungsmaß MAD (median absolute deviation), das den Median der absoluten Bioaktivitäts-Abweichungen (Verwendung von pK_i/pIC_{50}-Werten) von dem Knoten-Median angibt. Zur Charakterisierung von Ausreißern (typisch für SAR Hotspots und ACs) wurde zusätzlich die maximale absolute Abweichung (Abk. „MAXAD“) vom Knoten-Median für alle Knoten berechnet und der Median aller Werte (wie auch bei dem MAD) bestimmt. Zur Charakterisierung der SAR-(Dis-)Kontinuität in den Netzwerken wird der SARdisco Score (vgl. Kapitel 16.3) verwendet.

15. Vergleich der nächster Nachbarn (*k*NN-Regression)

15.1. Zielsetzung und Prinzip des Verfahrens der *k*NN

InSARa-Netzwerke verwenden eine andere Art der Molekülrepräsentation und Ähnlichkeitsmetrik als Fingerprint-basierte Ansätze. Nach dem SPP sollten bei adäquater Erfassung von Ähnlichkeit, d.h. sinnvolle molekulare Repräsentation und Ähnlichkeitsmaß, ähnliche Moleküle auch ähnliche biologische Aktivität aufweisen. Somit sollte auch eine Vorhersage von Molekülen basierend auf dem nachfolgend beschriebenen Prinzip der *k*-Nächsten-Nachbarn (engl. *k*-Nearest-Neighbor, Abk. *k*NN) umso erfolgreicher sein, desto besser die Ähnlichkeit von der Methode erfasst wird.

Der *k*-Nächsten-Nachbarn Algorithmus stammt aus der Muster-Erkennung und ist ein nicht-parametrischer Klassifikationsalgorithmus mit dem auf einfache Weise neue Objekte basierend auf der Mehrheitsentscheidung der *k*-nächsten-Nachbarn klassifiziert werden können.[403] Die Nähe der Nachbarn wird mit einem angemessen Distanz- oder Ähnlichkeitsmaß gemessen.[403] Das Prinzip kann aber auch zur Regression eingesetzt werden.[404] Da das Verfahren so einfach ist (kein besonderes Training notwendig), wird es oft auch „faules Lernen“ genannt.[403]

Auch im Bereich der QSAR wurde der *k*NN-Algorithmus von ZHENG und TROPSHA als einfache, nicht-lineare QSAR-Methode eingeführt.[405] In der ursprünglichen Methode wurde die biologische Aktivität eines Moleküls basierend auf dem arithmetischen Mittelwert seiner *k*NN im Trainingsdatensatz vorhergesagt.[405] SHEN et al. konnten jedoch eine Verbesserung der Ergebnisse durch Verwendung der gewichteten molekularen Ähnlichkeit zeigen.[406] Denn im Allgemeinen weist ein Molekül im chemischen Raum eine unterschiedliche Distanz bzw. Ähnlichkeit zu seinen nächsten Nachbarn auf.[406] Nähere bzw. ähnlichere Moleküle haben dabei eine höhere Wahrscheinlichkeit eine ähnliche Aktivität aufzuweisen, sodass ihnen ein höheres Gewicht zugeordnet wird.[406] Das Prinzip wird in Abbildung 15.1 zusammengefasst.

Zum Vergleich der inSARa-Ähnlichkeit mit Fingerprint-basierter Ähnlichkeit wurde daher ein auf dem oben beschriebenen Prinzip der *k*NN-Regression basierendes Verfahren entwickelt, das die Vorhersage von Bioaktivitäten mittels unüberwacht aufgebauter inSARa-Netzwerke ermöglicht. Da hierfür die Auswahl nächster Nachbarn entscheidend für die Güte der Vorhersage ist, kann über diese Vorhersagen indirekt ein Vergleich der Güte der Repräsentation und Ähnlichkeitserfassung erfolgen. Dies stellt das Ziel dieser Analyse dar. Es ist zu betonen, dass dieses Verfahren nicht darauf optimiert ist, gute Vorhersagen zu erzielen, die vergleichbar mit Techniken überwachten maschinellen Lernens (MLR, PLS, SVM, Random Forests u.v.m.) sind. Denn bei inSARa findet weder eine Modellselektion noch ein Training eines Modells statt, sondern die Netzwerke werden unüberwacht, ohne Berücksichtigung von Bioaktivitätsinformation basierend allein auf molekularer Ähnlichkeit aufgebaut.

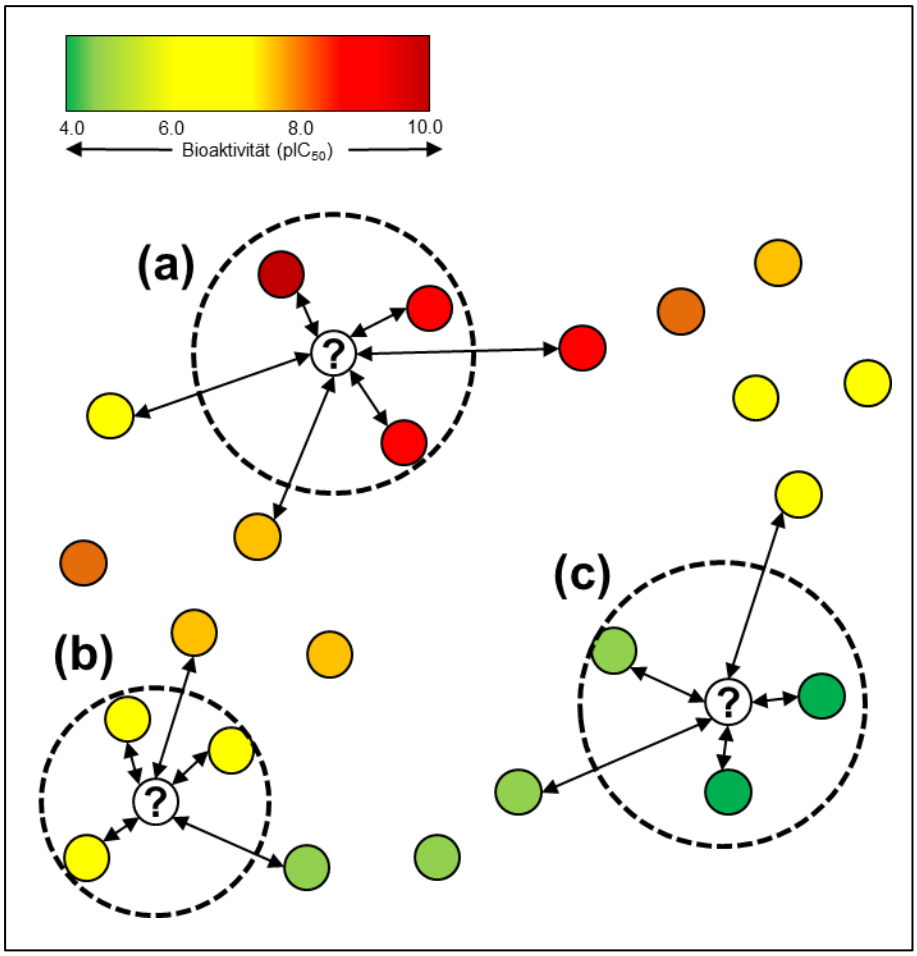


Abbildung 15.1. Prinzip der Bioaktivitäts-Vorhersage mittels *k*NN-Regression. Für *k*=3 werden die drei strukturell ähnlichsten nächsten Nachbarn aus dem Trainingsdatensatz (Moleküle mit bekannter Bioaktivität = bunte Knoten; Farbe entspricht Bioaktivität; gestrichelter Radius entspricht Entfernung des am drittweitesten entfernten nächsten Nachbarn vom Test-Objekt) für die Vorhersage der neuen Test-Moleküle (weiße Knoten mit Fragezeichen) berücksichtigt. Basierend auf diesem Prinzip würde für Test-Molekül (a) eine hohe, für (b) eine mittlere und für (c) eine sehr niedrige Bioaktivität vorhergesagt werden. Die beidseitigen Pfeile deuten die Entfernung im chemischen Raum bei einer bestimmten molekularen Repräsentation an.

15.2. Methode

a) inSARa-basierte *k*NN-Regression

In Abbildung 15.2 ist das Prinzip der auf *k*NN-Regression basierenden Vorhersage von biologischen Aktivitäten mittels InSARa-Netzwerks dargestellt. Im ersten Schritt werden Moleküle des Trainingsdatensatzes („Trainings-Molekül“) dazu verwendet ein inSARa-Netzwerk zu erzeugen. Hierzu werden die Moleküle wie in Kapitel 10 beschrieben zunächst in RGs umgewandelt, dann paarweise der MCS bestimmt und basierend hierdrauf eine hierarchische Netzwerk-Struktur unter Verwendung der Standardeinstellungen (Abbruchkriterium = 2% nicht-repräsentierte Moleküle, Mindest-MCS-Größe = 3, Ausschlussliste = aktiv) aufgebaut. Anschließend wird für jedes Molekül des Test-Datensatzes („Test-Molekül“) nach Codierung als RG der größte Substruktur-MCS des RGs des Moleküls bestimmt. Das ist der MCS-Knoten im Netzwerk, wo das Molekül

„steckenbleibt“. Die zugehörigen Moleküle sind hierbei die potentiell nächsten Nachbarn. Sollten mehr als *k*-nächste-Nachbarn verfügbar sein, werden als *k* NN diejenigen ausgewählt, die den größten RASCAL-Score aufweisen (vgl. Kapitel 2.3.3). Sollten zu wenig NN an dem oder den Knoten des „Steckenbleibens“ gefunden werden, wird im Netzwerk ein Schritt zurückgegangen und die Vorgängerknoten, die einen kleineren MCS repräsentieren mitberücksichtigt (vgl. Abbildung 15.2). Sollte trotzdem keine ausreichende Anzahl an *k* NN gefunden werden oder kein MCS-Knoten im Netzwerk existieren, der eine Substruktur des RGs des Test-Moleküls repräsentiert, wird das Molekül nicht vorhergesagt. Der vorhergesagte Bioaktivitätswert ist der gewichtete Mittelwert, der mit der RASCAL-Score-basierten Ähnlichkeit gewichtet wurde. Zudem wurde ein mehrfaches Gewicht für einen nächsten Nachbarn verwendet, sofern er an mehreren MCS-Knoten, wo das Test-Molekül im Netzwerk „steckenbleibt“ auftaucht. Hiermit soll der Tatsache Rechnung getragen werden, dass dieser NN eine höhere Wahrscheinlichkeit hat dem Test-Molekül ähnlich zu sein als andere NN.

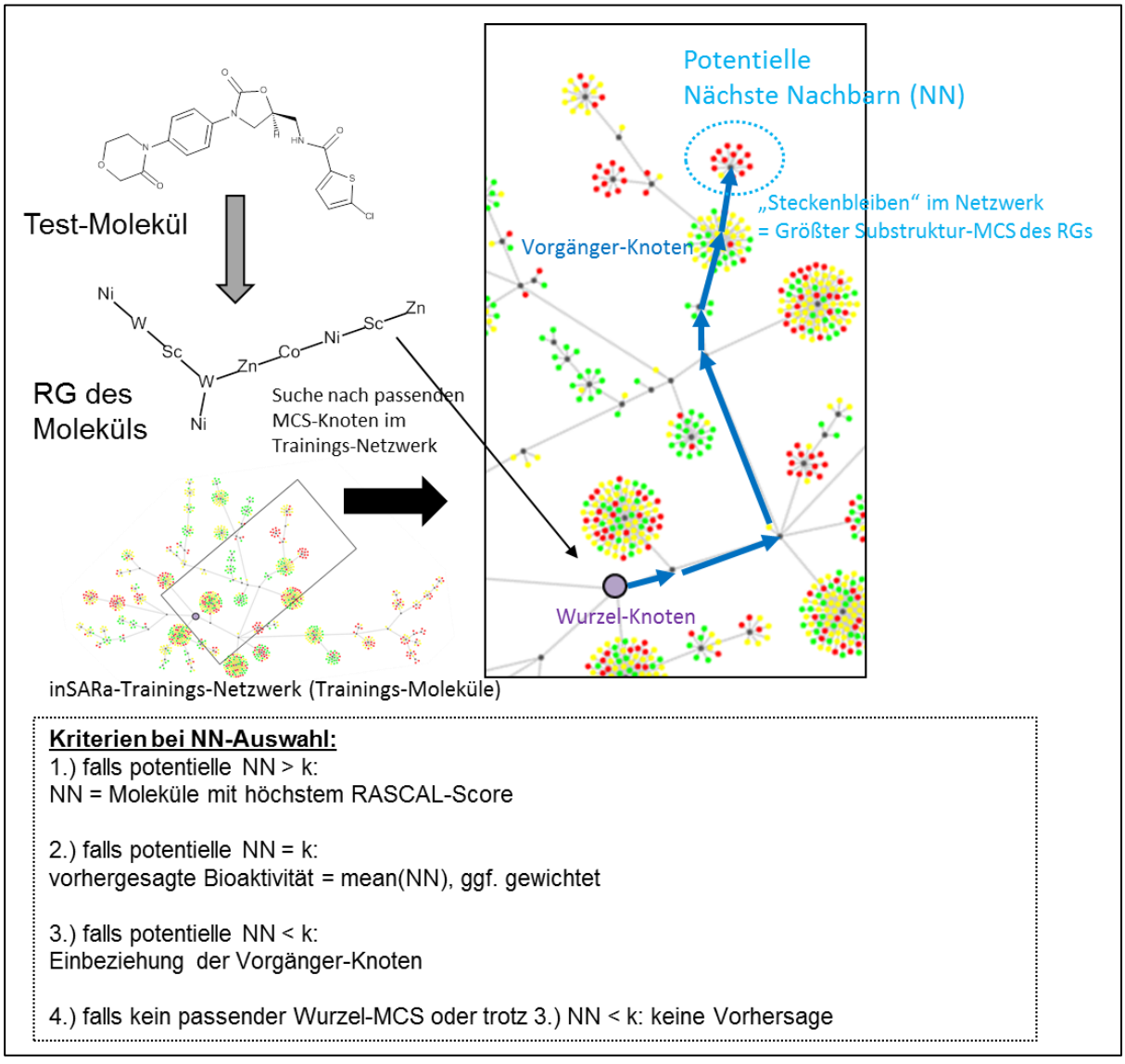


Abbildung 15.2. Prinzip der Vorhersage von Bioaktivitäten mittels inSARa-Netzwerk basierend auf *k*NN-Regression. Details siehe Text.

b) Fingerprint-basierte kNN-Regression

Bei der Fingerprint-basierten Vorhersage von Bioaktivitäten wurden zunächst für alle Moleküle des Test- und Trainingsdatensatzes die entsprechenden Fingerprints (für Details zu diesen siehe Kapitel 2.3.2 und 5.4) berechnet.

Als Fingerprints wurden die häufig verwendeten Substruktur-Fingerprints MACCS Keys in der binären (Abk. MACCS, 166 Bits, Codierung nur der An- und Abwesenheit von bestimmten Substrukturen) und der Integer-Variante (Abk. MACCSF, 166 Bits, Codierung der Häufigkeit bestimmter Substrukturen), sowie der zirkuläre ECFP-Fingerprint mit einem Durchmesser von 4 Bindungen (Abk. ECFP4), der Pfad-basierte Open Babel Fingerprint FP2 (Abk. FP2, 1024 Bits) und der topologische Pharmakophor-Fingerprint CATS2D (Abk. CATS2D, maximale Pfadlänge 10) ausgewählt. MACCS wurde mit MOE[188] berechnet, für MACCSF und ECFP4 wurden zusätzliche MOE-SVL-Skripte[402, 407] verwendet, der FP2 wurde mit Open Babel[115] berechnet und für CATS2D wurde eine Python-Implementierung von GUHA[408] verwendet.

Zusätzlich wurden noch drei simple RG-Fingerprints implementiert. Der einfachste RG-Fingerprint „RG_atom_count“ ist ein 14-Bit-FP (Anzahl an RG-Pseudoatomen) und zählt nur die Häufigkeit des Vorkommens eines bestimmten RG-Atomtyps ohne Berücksichtigung von Konnektivitäten. Der zweite RG-Fingerprint „RG_atompairs“ ist ein 650-Bit-FP und erfasst alle vorhandenen Atompaare innerhalb einer bestimmten Distanz. Bei 14 verschiedenen Atomtypen gibt es 105 Atompaar-Kombinationen. Mittels Distanzmatrix wird die Entfernung zwischen den einzelnen Atompaaren bestimmt. Es werden nur Pfade von 1 bis 6 Bindungen codiert, längere Pfade werden so behandelt wie ein Pfad der Länge 6 (analog zu HARPER et al.[201]). Auf weitere spezielle Codierungen wird verzichtet (Unterschied zu HARPER et al.[201]). Der dritte RG-Fingerprint „RG_atompairs_fuzzy“ lässt in Anlehnung an STIEFL et al.[239] eine gewisse Unschärfe bezüglich der Distanz, die zwei Atome trennt, zu. So werden die zwei benachbarten Distanzbins ebenfalls um den „fuzzy-Faktor“ 0.3 (statt 1.0 für die tatsächliche Distanz) inkrementiert.

Das Prinzip der Fingerprint-basierten Vorhersagen ist in Abbildung 15.3 dargestellt. Für ein Test-Molekül wird jeweils die Tc-Ähnlichkeit zu allen Training-Molekülen berechnet. Als *k* NN werden die Moleküle mit dem höchsten Tc verwendet. Bei mehreren Molekülen mit gleichem Tc werden aus dieser Menge zufällig so viele Moleküle ausgewählt bis *k* NN gefunden sind. Die vorhergesagte Bioaktivität ist der gewichtete Mittelwert der *k*NN, der mit der jeweiligen Tc-Ähnlichkeit gewichtet wird.

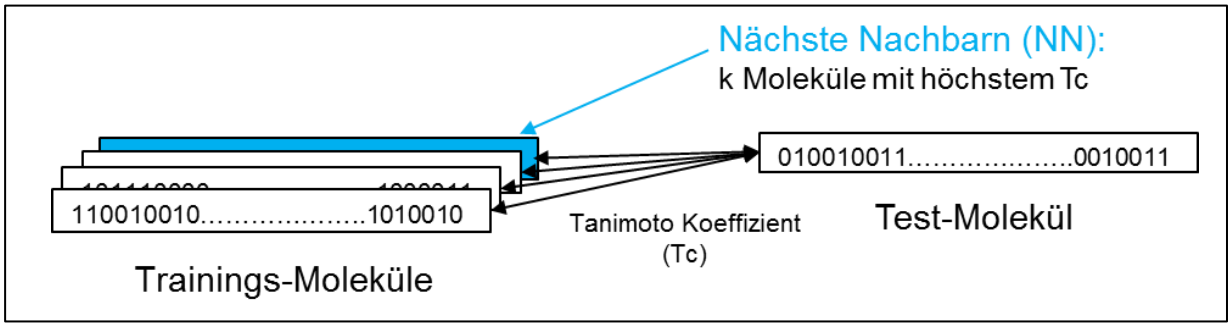


Abbildung 15.3. Fingerprint-basierte KNN-Regression.

15.3. Verwendete Datensätze

Für die Vorhersagen wurden die sechs in Tabelle 11.1 aufgeführten Datensätze aus der BindingDB verwendet. Da die Bioaktivitäts-Verteilung in den Datensätzen nicht gleichmäßig ist, sondern zumeist einer Gauß-Kurven-ähnliche Verteilung mit einem Maximum bei mittlerer Aktivität (pK_i oder pIC_{50} zwischen etwa 6 und 7) zu beobachten ist, wurden nicht alle Moleküle gleichzeitig für die Analyse verwendet. Stattdessen wurden jeweils zufällig gleichgroße Aktivitätsklassen aus der Gesamtmenge an Molekülen gezogen („undersampling“ der größeren Aktivitätsklassen). Dieser ausgeglichene, neu zusammengestellte Datensatz wird dann zufällig in Trainings- und Testdaten im Verhältnis 2/3 und 1/3 aufgeteilt. Dieses Ziehen eines ausgeglichenen Datensatzes und anschließendes Aufspalten in Test- und Trainingsdaten wurde für jede Zielstruktur 500-mal wiederholt, um eine zuverlässige Aussage über den Streuungsbereich der Ergebnisse zu erhalten. Es ist zu beachten, dass Moleküle aus der aus der im ursprünglichen Datensatz geringer besetzten Aktivitätsklasse häufiger vorhergesagt werden als Moleküle aus der dichter besetzten Aktivitätsklasse. Für die einzelnen Targets wurden zuvor jeweils Klassengrenzen für hohe, mittlere und schwache Bioaktivität festgelegt (vgl. Abbildung 15.4) Ebenfalls wurde eine Maximalaktivitätsgrenze festgelegt, um sehr selten auftretende sehr hoch aktive Moleküle, die aufgrund von fehlenden Nachbarn im Bioaktivitätsraum nur schwer vorhersagbar sind, von den Vorhersagen auszuschließen. Anschließend wurde analysiert, wie viele Moleküle des Gesamtdatensatzes den jeweiligen Klassen zugeordnet werden können. Ausgehend von der kleinsten Klasse wurde dann individuell die Größe der Aktivitätsklassen im zusammengestellten Datensatz festgelegt (vgl. Tabelle 15.1).

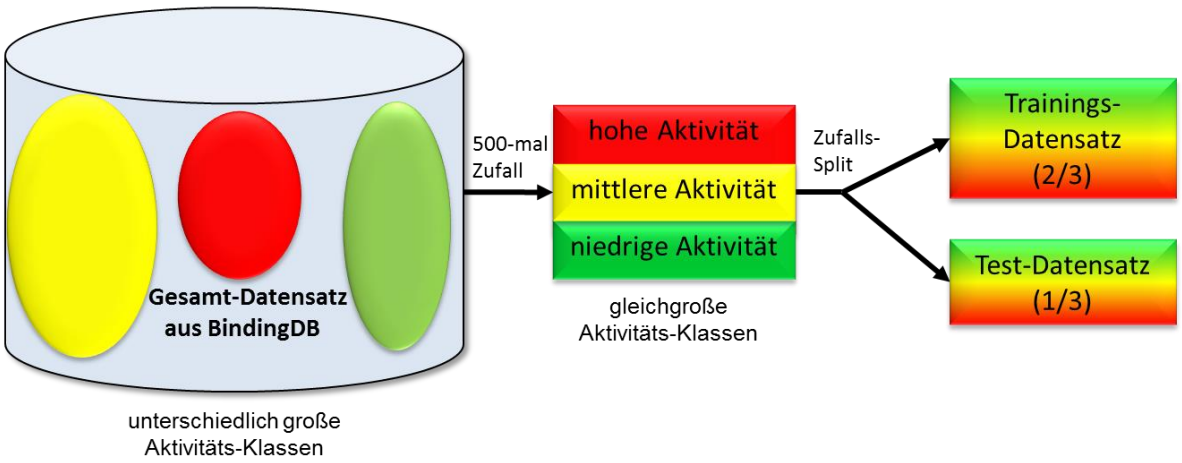


Abbildung 15.4. Zusammenstellung der Test- und Trainingsdatensätze aus den ursprünglichen BindingDB-Datensätzen. Details siehe Text.

15.4. Auswertung

Zur Auswertung wurden für jeden der 500 Test-Datensätze jeweils die folgenden gebräuchlichen Gütekriterien[409] nach den nachfolgend aufgeführten Formeln berechnet (zur Veranschaulichung siehe Abbildung 15.5).

Die *Quadratwurzel des mittleren quadrierten Fehlers der Datenvorhersage (Abk. RMSEP, engl. Root Mean Squared Error of Prediction)* wird als absolutes Fehlermaß verwendet. Der RMSEP berechnet sich aus der Quadratwurzel aus dem *vorhergesagten Fehler der Summe der Abweichungsquadrate (Abk. PRESS, engl.: PRedicted Error Sum of Squares)* dividiert durch die Anzahl an vorhergesagten Testobjekten n_{Test}. $y_{i,Test}$ ist hierbei der experimentell ermittelte Bioaktivitätswert des Test-Moleküls$_i$ und $\hat{y}_{i,Test}$ der vorhergesagte Bioaktivitätswert für das Test-Molekül *i*.

$$\mathrm{RMSEP_{Test}} = \sqrt{\frac{\mathrm{PRESS_{Test}}}{\mathrm{n_{Test}}}} \qquad \textbf{(15.1)}$$

$$\mathrm{PRESS_{Test}} = \sum_{i=1}^{n_{Test}} (y_{i,Test} - \hat{y}_{i,Test})^2 \qquad \textbf{(15.2)}$$

Als relative Gütekennzahl wird der *quadrierte, multiple Korrelationskoeffizient der Datenvorhersage R^2 (Bestimmtheitsmaß)* berechnet. Für $\bar{y}_{Training}$ wird der Mittelwert aller Trainings-Moleküle eingesetzt.

$$\mathrm{R^2_{Test}} = 1 - \frac{\mathrm{PRESS_{Test}}}{\sum_{i=1}^{n_{Test}} (y_{i,Test} - \bar{y}_{Training})^2} \qquad \textbf{(15.3)}$$

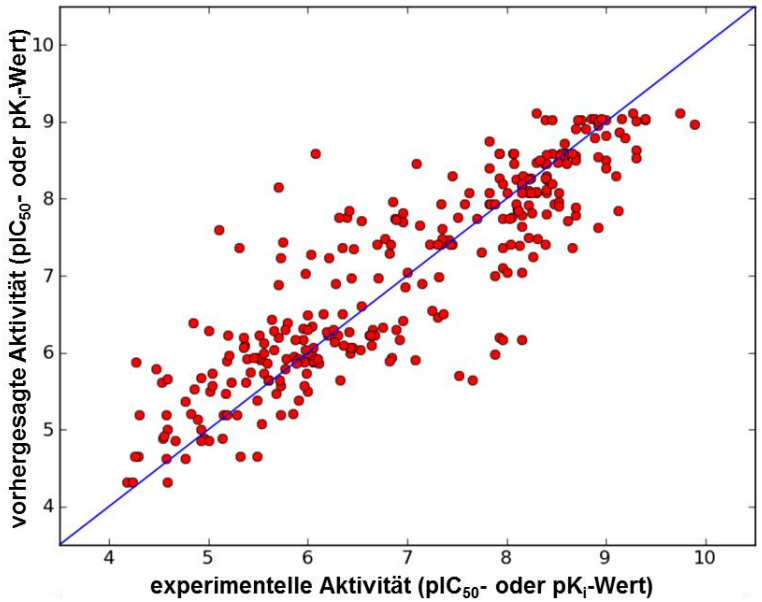


Abbildung 15.5. Auswertung der *k*NN-basierten Bioaktivitätsvorhersage für einen Test-Trainings-Datensatz-Split. Im Idealfall (ohne experimentellen Fehler und bei perfekter molekularer Repräsentation) würde man erwarten, dass alle roten Punkte auf der blauen Linie (vorhergesagte Aktivität = experimentell bestimmte Aktivität) lägen. In diesem Fall wäre der R^2 = 1 und RMSEP = 0. In diesem Beispiel ist der R^2 = 0.79 und RMSEP = 0.66.

Tabelle 15.1. Übersicht über Charakteristika der Datensätze für die *k*NN-basierte Bioaktivitäts-Vorhersage. Für weitere Details bezüglich der Datensätze siehe Tabelle 11.1 (in Kapitel 11.2) und Tabelle 26.2 (im Anhang).

Ziel-struktur (Abkür-zung)	**Anzahl an Molekülen nach Vorbereitung und RG-Erzeugung**	**Anzahl an Molekülen für Vorhersage (Aufteilung: 2/3 Trainings-, 1/3 Test-datensatz)**	**Bio-aktivitäts-Wert**	**Bioaktivitäts-Verteilung**			**Aktivitätsklassen-Grenzen (Moleküle pro Klasse)**			**Anzahl der Moleküle in der Zufalls-Stichprobe**		
				Min	**Max**	**Mean**	**Niedrige Aktivität (L)**	**Mittlere Aktivität (I)**	**Hohe Aktivität (H)**	**L-Klasse**	**I-Klasse**	**H-Klasse**
FXA	1736	900	pIC_{50}	4.07	10.70	7.13	≤ 6 (367)	>6 - <8 (858)	≥8 - <10 (507)	300	300	300
COX2	2349	1170	pIC_{50}	4.01	11.22	6.33	≤ 6 (960)	>6 - <7.5 (974)	≥7.5 - <10 (408)	390	390	390
CB1	1957	1080	pK_i	4.10	9.96	6.96	≤ 6 (440)	>6 - <8 (1140)	≥8 - <10 (377)	360	360	360
CDK2	1675	765	pIC_{50}	4.01	9.52	6.72	≤ 6.5 (632)	>6.5 - <8 (665)	≥8 - <10 (270)	255	255	255
P38	2446	1440	pIC_{50}	4.06	10.22	7.04	≤ 6.5 (723)	>6.5 - <8 (1224)	≥8 - <10 (494)	480	480	480
THR	2852	1530	pK_i	4.00	12.19	6.76	≤ 6 (1024)	>6 - <8 (1204)	≥8 - <10 (548)	510	510	510

16. Anwendung: SAR-Interpretation

16.1. Regeln für die interaktive SAR-Analyse

a) Lokale SAR-Analyse (Fokussierung auf einzelne MCS-Knoten)

Bei der Analyse von großen Datensätzen (mehrere hundert bis tausend Moleküle), können die resultierenden inSARa-Netzwerke groß werden. Aus diesem Grund wurden einfache Regeln abgeleitet, die das Auffinden interessanter SARs in den inSARa-Netzwerken erleichtern.

Die Analyse sollte sich zunächst auf die terminalen Knoten konzentrieren, da die RG-MCSs spezifischer sind als die MCSs an nicht-terminalen Knoten. Die zugehörigen Moleküle sind daher auch strukturell ähnlicher und die SARs somit einfacher zu interpretieren. Empirische Analysen haben gezeigt, dass MCS Knoten, deren zugehörige MCSs eine Größe von mehr als 8 oder 9 Pseudoatomen aufweisen, näher betrachtet werden sollten. Es hat sich hierbei herausgestellt, dass die zugehörigen Moleküle in den meisten Fällen als chemisch „ähnlich“ erkannt werden. An diesen Knoten sind also oftmals interpretierbare Beziehungen zu finden und SAR Trends können abgeleitet werden. Bei einzelnen Datensätzen ist diese Daumenregel aufgrund sehr großer RG-Größen anzupassen (vgl. THR in Abschnitt 21.1).

Zusätzlich sollte man sich in Abhängigkeit der der SAR-Analyse zugrunde liegenden Fragestellung auf verschiedene MCS-Knotentypen konzentrieren. Es lassen sich drei wichtige Typen unterscheiden, die sich deutlich in der vorzufindenden SAR-Informationen unterscheiden. Ein Beispiel für jeden dieser Knoten-Typen ist in Abbildung 16.1. zu finden.

1.) Wenn man MCS-Knoten betrachtet, die *gleichzeitig mit Molekül-Knoten verknüpft sind, deren Mehrheit rot gefärbt ist und eine kleine Anzahl ist grün gefärbt oder umgekehrt* (d.h. die zugehörigen Moleküle zeigen große Unterschiede in der Bioaktivität), können *sprunghafte SARs* oder *nicht-bioisostere Austausche* identifiziert werden. Im Gegensatz zu Fingerprint-basierten Definitionen von sprunghaften SARs muss hierbei kein von dem Fingerprint-Typ abhängiger Schwellenwert definiert werden (vgl. Abschnitt 2.5.1).

2.) Wenn der medizinische Chemiker an der Identifizierung von *bioisosteren Austauschen* (vgl. Abschnitt 2.5.2) interessiert ist (z.B. in der Leitstruktur-Optimierung), dann sollten MCS-Knoten betrachtet werden, die nur mit *einfarbigen Molekülknoten* verbunden sind (d.h. die zugehörigen Moleküle weisen ähnliche Bioaktivitäten auf).

3.) Ein weiterer interessanter Knotentyp in inSARa-Netzwerken sind MCS-Knoten, die mit Molekül-Knoten verbunden sind, die *verschiedene Farben* haben (d.h. die zugehörigen Moleküle zeigen eine hohe Varianz bezüglich der Bioaktivität). Im Gegensatz zum ersten Knotentyp findet sich in der Regel keine Mehrheit von einer Farbe, jedoch eine hohe Varianz bezüglich der Farbe bzw. der Bioaktivität. An diesen *SAR Hotspots* können zumeist molekulare Eigenschaften identifiziert werden, die die Bioaktivität entscheidend beeinflussen (vgl. Kapitel 2.5.3). Dies ist von besonderem Interesse in der Leitstruktur-Optimierung.

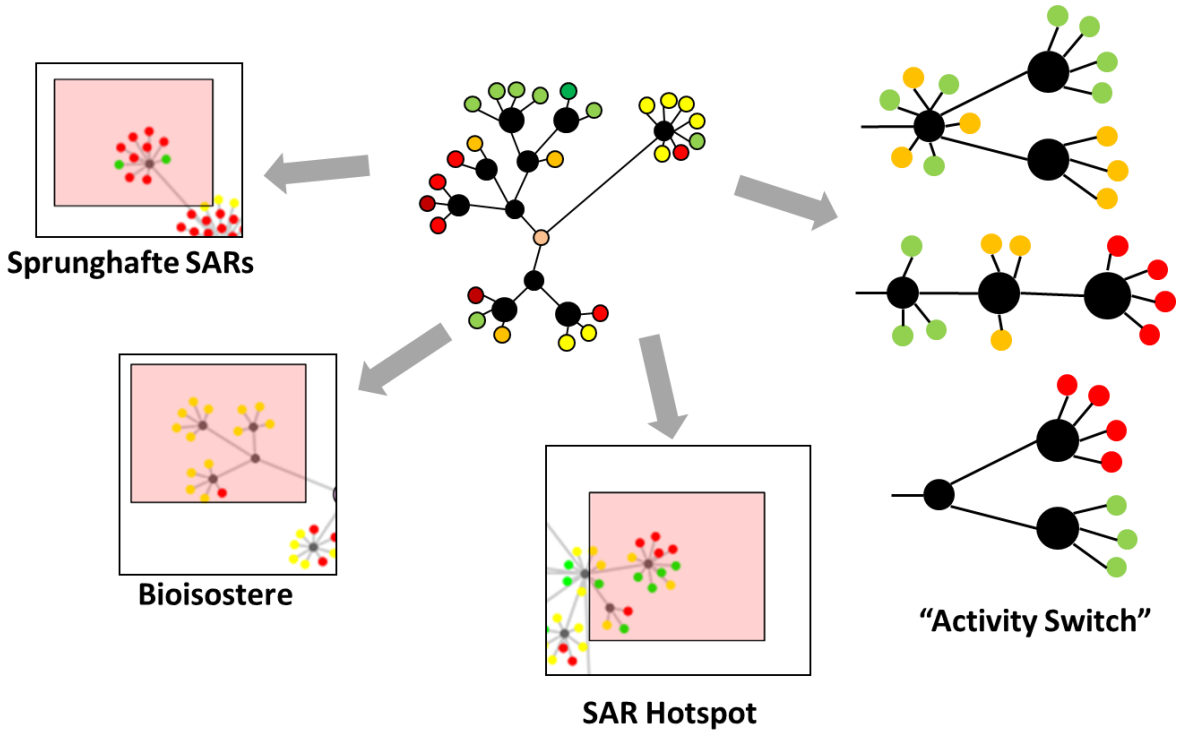


Abbildung 16.1. Erkennen verschiedener SAR-Informationen in inSARa-Netzwerken.

b) Subnetzwerk-SAR-Analyse (Berücksichtigung benachbarter MCS-Knoten)

Beim Betrachten nicht nur einzelner Knoten, sondern ganzer Netzwerkbereiche, insbesondere benachbarter MCS-Knoten, können Pfade erkannt werden, die im Folgenden als *“Activity Switches”* bezeichnet werden sollen. Ein “Activity Switch” wird im Folgenden definiert als ein Ast im inSARa-Netzwerk, in dem sich die Bioaktivität trendmäßig von einem zum anderen MCS-Knoten verändert (z.B. von niedrig zu hoch aktiv oder umgekehrt). Diese Pfade sind wichtig für das Erkennen von bioaktivitätsbestimmenden pharmakophoren Eigenschaften. Verschiedene Beispiele sind modellhaft dargestellt in Abbildung 16.1 zu finden.
Anmerkung: Unabhängig von inSARa definierten MEDINA-FRANCO et al. im Juni 2013 „Activity Switches“ als „spezifische Substitutionen, die gegensätzliche Auswirkungen auf die Bioaktivität von Molekülen an zwei Zielstrukturen haben“[410]. MEDINA-FRANCO et al. umschreiben mit dem Begriff also inverses SAR-Verhalten an *verschiedenen* Targets, während im Zusammenhang mit inSARa pharmakophore Merkmale gemeint sind, die systematisch die Bioaktivität an *einer* Zielstruktur beeinflussen.

16.2. Automatisierte SAR-Analyse: inSARa^{auto}

Motivation und Prinzip

Die zuvor vorgestellte Form der SAR-Interpretation mittels interaktiver Netzwerk-Navigation und manueller Auswahl von interessanten MCS-Knoten ist bei großen Netzwerken zeitaufwändig und z.T. auch stark subjektiv beeinflusst. Zudem ist ein schneller Vergleich von einer Vielzahl von Datensätzen bezüglich der enthaltenen Menge und Art von SAR-Information nur schwer möglich.

Um das Erkennen von interessanter, leicht interpretierbarer SAR-Information in den großen Netzwerken zu vereinfachen bzw. zu beschleunigen, wurde daher das inSARa^{auto} Verfahren entwickelt, das die zusätzliche automatisierte Analyse der MCS-Knoten eines inSARa-Netzwerkes ermöglicht. Hierbei werden die Knoten bezüglich struktureller Ähnlichkeit und Variabilität in der Bioaktivität untersucht. Dafür wurden Maßzahlen zur Charakterisierung dieser Knoten-Eigenschaften definiert. Dabei ist es wichtig, dass diese Maßzahlen die in Abschnitt 16.1a beschriebenen Merkmale der drei MCS-Knoten-Typen erfassen und zur Abgrenzung der Knoten-Typen geeignet sind. MCS-Knoten bekannter SAR-Information wurden anschließend mit den ausgewählten Maßzahlen analysiert. Basierend darauf wurden Grenzen zur Unterscheidung der verschiedenen MCS-Knoten bezüglich der repräsentierten SAR-Information festgelegt und abermals an verschiedenen Datensätzen mittels manueller visueller Inspektion der extrahierten MCS-Knoten validiert. Im Anschluss wurden die definierten Grenzwerte abermals angepasst.

Im Folgenden werden die ausgewählten Maßzahlen vorgestellt und die Überlegungen zur Abgrenzung verschiedener Knoten-Typen durch diese aufgezeigt, sowie die sich aus der iterativen manuellen Optimierung der Grenzwerte resultierenden Werte gezeigt.

Auswahl von Maßzahlen

a) Charakterisierung struktureller MCS-Ähnlichkeit

Für erfolgreiche SAR-Interpretation ist eine bestimmte strukturelle Mindest-Ähnlichkeit eine wichtige Voraussetzung. In Abschnitt 16.1 wurde betont, dass einfach interpretierbare SARs v.a. an den terminalen Knoten bzw. Knoten mit größerem MCS gefunden werden. Wichtig ist es, die MCS-Größe immer im Verhältnis zur Größe der zugehörigen RGs zu betrachten (analog beispielsweise zu dem in Kapitel 2.3.3 vorgestellten RASCAL-Score zur paarweisen MCS-basierten Ähnlichkeits-Bestimmung). Diesem Prinzip folgend wurde daher der *MCS-Sim-Score* entwickelt, um die MCS-Ähnlichkeit an einem bestimmten MCS-Knoten bestimmen zu können. Hierdurch ist es möglich eine für die Interpretierbarkeit wichtige Mindest-MCS-Ähnlichkeit an den aufgrund bestimmter Bioaktivitätsmuster potentiell interessanten MCS-Knoten sicherzustellen.

Hohe strukturelle Ähnlichkeit kann angenommen werden, wenn der RG-MCS im Vergleich zur durchschnittlichen RG-Größe groß ist. Für einen MCS-Knoten *i* im Netzwerk mit *n* zugehörigen Molekülen wird die strukturelle Ähnlichkeit *MCS-Sim-Score* daher wie folgt berechnet:

$$\mathrm{MCS-Sim-Score(i)} = \frac{\text{Größe des RG} - \mathrm{MCS_i}}{\left(\sum_{j=1}^{n} \text{RG} - \text{Größe des Molekül}_j\right)/n} \qquad (16.1)$$

b) Analyse der Bioaktivitäts-Variabilität

Unter der Voraussetzung einer definierten strukturellen Mindest-Ähnlichkeit ist (wie man anhand von Abbildung 16.1 sieht) das Bioaktivitätsmuster bzw. die Variabilität in der Bioaktivität (visuell durch ein bestimmtes Farbmuster erkennbar) an einem MCS-Knoten entscheidend für die potentielle Zuordnung zu einem der drei SAR-Typen.

Variabilität in der Bioaktivität lässt sich unterschiedlich messen. Umfassende Analysen haben jedoch gezeigt, dass es schwierig ist eine einzelne Maßzahl zu finden, die zur Differenzierung der verschiedenen SAR-Phänomene geeignet ist. Stattdessen hat sich die Kombination der nachfolgenden Maßzahlen als vielversprechend für die erfolgreiche „Farbmustererkennung“ herausgestellt.

- **Standardabweichung (Abk. SD):**
  Die Standardabweichung der Bioaktivitäten aller Moleküle an einem MCS-Knoten drückt die mittlere Abweichung der Bioaktivitäten vom Mittelwert an diesem Knoten aus. Einzelne Ausreißer mit stark abweichender Bioaktivität (wie z.B. bei sprunghaften SARs) lassen sich hierbei nur schwer erkennen. Geringe Standardabweichungen deuten allgemein darauf hin, dass die Moleküle bezüglich der Bioaktivität sehr homogen sind (bioaktivitätserhaltende Modifikationen bzw. bioisosterer Austausch). Eine große SD ist typisch für SAR Hotspots.

- **Differenz zwischen Minimal- und Maximal-Wert (Abk. ΔMinMax):**
  Es wird die Differenz zwischen den Molekülen am MCS-Knoten mit dem kleinsten und größten Bioaktivitätswert berechnet. Kleine Differenzen deuten homogenes Verhalten an („Bioisosterie“), große Differenzen drücken große Gesamt-Streuung (AC oder SAR Hotspot) aus. Da jedoch nur Extremwerte betrachtet werden, ist keine Aussage möglich, ob generell eine hohe Bioaktivitäts-Variabilität an diesem Knoten vorliegt (wie beim SAR Hotspot) oder ob es sich bei dem Minimal- und Maximalwert um einen Ausreißer handelt, sich aber die meisten Moleküle nur unwesentlich bezüglich der Bioaktivität unterscheiden (wie beim AC).

- **Maximale Abweichung vom Median (Abk. ΔMaxMedian):**
  Es wird für alle Moleküle am MCS-Knoten der Median der Bioaktivität berechnet und dann die maximal auftretende Differenz zu diesem Wert bestimmt. Dieses Kriterium wird zur Unterscheidung von sprunghaften SARs (Wert entspricht in etwa ΔMinMax) und SAR Hotspots (Wert deutlich kleiner als ΔMinMax) verwendet.

Qualitative Charakterisierung verschiedener inSARa-Knoten-Typen

Tabelle 16.1 fasst die Überlegungen zusammen, wie mit Hilfe der ausgewählten Maßzahlen die verschiedenen Typen an SAR-Information unterschieden werden können.

Tabelle 16.1. Qualitative Kriterien zur Charakterisierung der verschiedenenen MCS-Knoten-Typen mittels inSARa^{auto}. Für die Definition der einzelnen Maßzahlen und weitere Details siehe Text.

Kriterium		**Hinweis auf**		
		SAR Hotspot	**sprunghafte SARs**	**Bioisosterie oder bioaktivitätserhaltende Modifikationen**
Variabilität in Bioaktivität	**SD**	hoch	eher niedrig	niedrig
	ΔMinMax	hoch	hoch	niedrig
	ΔMaxMedian	≠ ΔMinMax	≈ ΔMinMax	niedrig
Strukturelle Ähnlichkeit (MCS-Sim-Score)		hoch	(sehr) hoch	(sehr) hoch
Median der Bioaktivitäten		keine Vorgabe		mittel bis hoch

Das Erkennen von Bioisosterie bzw. bioaktivitätserhaltenden strukturellen Modifikationen ist am einfachsten aufgrund der geforderten geringen Bioaktivitäts-Variabilität (bei allen 3 Kriterien werden geringe Werte gefordert) und der deutlichen Abgrenzung von den anderen beiden MCS-Knoten-Typen. Da bioisosterer Austausch zur Leitstruktur-Optimierung dienen soll, sind nur MCS-Knoten mit einer bestimmten Mindest-Bioaktivität (mittlere oder hohe biologische Aktivität = gelbe oder rote Knoten) von Relevanz. Daher wird als zusätzliches Kriterium der Median der Bioaktivitäten aller Moleküle eines MCS-Knoten berechnet.

Das Erkennen von SAR Hotspots und sprunghaften SARs bzw. deren Differenzierung ist deutlich schwieriger. Gemeinsam haben beide, dass die Differenz zwischen dem Molekül mit der niedrigsten und höchsten Bioaktivität ΔMinMax sehr groß ist (im Gegensatz zur Bioisosterie). Bei echten sprunghaften SARs weisen jedoch die restlichen Moleküle eine ähnliche biologische Aktivität auf wie eines der beiden Extremwert-Moleküle, sodass die Abweichung vom Median-Wert ΔMaxMedian dem Wert von ΔMinMax sehr ähnlich ist. Beim SAR Hotspot liegt der Median-Wert im Idealfall in etwa in der Mitte der beiden Extramwerte. ΔMinMax und ΔMaxMedian sollten sich somit deutlich unterscheiden. Da ACs definiert sind als geringer struktureller Unterschied, der zu einer großen Bioaktivitätsveränderung führt, ist es sinnvoll, einen sehr hohen MCS-Sim-Score für sprunghafte SARs vorauszusetzen.

Quantitative Kriterien zur MCS-Knoten-Charakterisierung

In Tabelle 16.2 sind die optimierten Werte zur Charakterisierung der verschiedenen MCS-Knoten-Typen zusammengefasst.

Tabelle 16.2. Quantitative Kriterien zur automatischen MCS-Knoten-Charakterisierung mittels inSARa^{auto}. Für weitere Details siehe Text.

Kriterium	**MCS-Knoten erfüllt Charakterik für folgende SAR-Information:**		
	SAR Hotspot	**sprunghafte SARs**	**Bioisosterie oder bioaktivitäts-erhaltende Modifikationen**
SD	≥ 1.0	≥ 0.5	≤ 0.9
ΔMinMax	≥ 2.5	≥ 2.0	≤ 1.5
ΔMaxMedian	≥ 1.5	keine Vorgabe	≤ 0.9
|(ΔMinMax – ΔMaxMedian)|	> 0.5	a) Moleküle > 2: ≤ 0.5 b) Moleküle = 2: keine Vorgabe	keine Vorgabe
MCS-Sim-Score	≥ 0.6	≥ 0.8	≥ 0.8 (bzw. ≥ 0.6)
Mindest-Anzahl an Molekülen	3	2	2
Median(pIC_{50}/pK_i)	keine Vorgabe		pIC_{50}/pK_i ≥ 6.0

Anwendung: Analyse und Vergleich verschiedener Datensätze

Zur Anwendung wurden die in Kapitel 11.2 aufgeführten Datensätze bezüglich der enthaltenen SAR-Information analysiert unter Verwendung der folgenden Standardeinstellungen: Mindest-MCS-Größe = 5 RG-Atome, Abbruch-Kriterium = 2% nicht-repräsentierte Moleküle, Ausschlussliste = aktiv.

Durch diese automatisierte SAR-Analyse ist es möglich sehr schnell eine große Zahl an Datensätzen bezüglich ihrer Charakteristika zu vergleichen bzw. einen bestimmten Typ potentieller SAR-Information ohne subjektive, manuell relativ aufwändige Netzwerk-Analyse betrachten zu können. Zum einen ist es möglich MCS-Knoten eines bestimmten Typs inklusive zugehöriger Moleküle aus dem Netzwerk zu extrahieren oder aber im Netzwerk entsprechend hervorheben zu lassen. Diese Information kann ebenfalls dazu verwendet werden, um Netzwerke in ihrer Komplexität zu reduzieren. Hierfür würden alle Knoten abgesehen von den potentiell interessanten Knoten und ggf. Verknüpfungs-Knoten vom Netzwerk entfernt werden.

16.3. Globale automatisierte SAR-Charakterisierung: SARdisco Score

Zielsetzung

In Anlehnung an die in Kapitel 2.4.2 beschriebenen Maßzahlen zur globalen quantitativen SAR-Charakterisierung wurde auf der Basis des vorangehend beschriebenen inSARa^{auto}-Ansatzes der SARdisco Score (characterization of SAR discontinuity and continuity in inSARa-networks) entwickelt. Der SARdisco Score ist eine Maßzahl zur globalen Charakterisierung von inSARa-Netzwerken bezüglich SAR-(Dis-)Kontinuität. Es ermöglicht den schnellen Vergleich einer Vielzahl von Datensätzen oder die grobe Erfassung von Veränderungen in Netzwerken desselben Targets (z.B. nach Veränderung der RG-Definition oder anderer Einstellungen).

Methode

Zur Bestimmung des SARdisco Scores werden alle MCS-Knoten, die eine bestimmte Mindest-Ähnlichkeit aufweisen (Anwendbarkeit des SPP; hier: MCS-Sim-Score ≥ 0.6, vgl. Abschnitt 16.2), eingeteilt nach kontinuierlichem (dem SPP folgend: hohe strukturelle Ähnlichkeit, geringe Bioaktivitätsdifferenz) und diskontinuierliches SAR-Verhalten (entgegen dem SPP: hohe strukturelle Ähnlichkeit, große Bioaktivitätsdifferenz).

Zur Identifizierung von *SAR-Kontinuität* werden die für inSARa^{auto} beschriebenen Kriterien zur Identifizierung von bioaktivitätserhaltenden Modifikationen (vgl. Tabelle 16.2) verwendet. Da auch MCS-Knoten, die nur schwach aktive Moleküle repräsentieren, dem SPP folgen (=SAR-Kontinuität), wird im Gegensatz zu den Kriterien in Tabelle 16.2 auf die Definition einer Mindest-Bioaktivität verzichtet. Zur Identifizierung von *SAR-Diskontinuität* werden die für inSARa^{auto} beschriebenen Kriterien zur Identifizierung von ACs und SAR Hotspots (vgl. Tabelle 16.2) verwendet. Darüberhinausgehend sind in den Netzwerken zumeist weitere MCS-Knoten zu finden, die zwar die Mindest-MCS-Ähnlichkeit aufweisen, aber nach den definierten Kriterien ist keine eindeutige Klassifikation als kontinuierliches oder diskontinuierliches Verhalten möglich. In Anlehnung an Abschnitt 2.4.3 werden diese MCS-Knoten der Kategorie „*heterogenes SAR-Verhalten*“ zugeordnet.

Basierend auf den vorgenannten Definitionen zur MCS-Knoten-Klassifikation berechnet sich der SARdisco Score wie folgt:

$$\text{SARdisco Score} = \frac{\text{MCSs}_{\text{cont}} + \text{MCSs}_{\text{hetero}} \cdot 0.5}{\text{MCSs}_{0.6}} \qquad (16.2)$$

$\text{MCSs}_{\text{cont}}$ ist hierbei die Anzahl an MCS-Knoten, die der Definition von kontinuierlichem SAR-Verhalten entsprechen. $\text{MCSs}_{\text{hetero}}$ ist die Anzahl an MCS-Knoten, die heterogenes SAR-Verhalten zeigen. $\text{MCSs}_{0.6}$ ist die Gesamtzahl an MCS-Knoten, die eine MCS-Ähnlichkeit von 0.6 (MCS-Sim-Score) aufweisen, d.h. die Summe aus $\text{MCSs}_{\text{cont}}$, $\text{MCSs}_{\text{hetero}}$ und den MCS-Knoten, die der Definition von diskontinuierlichem SAR-Verhalten entsprechen. Da sich $\text{MCSs}_{\text{hetero}}$ partiell kontinuierlichem und diskontinuierlichem SAR-Verhalten zuordnen lässt, wird es mit dem Faktor 0.5 gewichtet. Auf dieser Basis kann der SARdisco Score analog

zum SAR-Index (vgl. Abschnitt 2.4.2) Werte von 0 (= Netzwerke sind nur durch SAR-Diskontinuität gekennzeichnet) bis 1 (= Netzwerke zeigen nur SAR-Kontinuität) annehmen. Werte um 0.5 zeigen heterogenes SAR-Verhalten (d.h. entweder Knoten, die sich nicht eindeutig klassifizieren lassen oder es liegt ein ausgeglichenes Verhältnis zwischen SAR-(Dis-)Kontinuität vor).

Anwendung

Zur Anwendung wurden die in Tabelle 17.2 aufgelisteten 140 Datensätze aus Abschnitt 17.5, die verschiedenste Targets repräsentieren, mittels SARdisco Score charakterisiert. Im Gegensatz zu der in Kapitel 17 beschriebenen Analyse wurde für die Zusammenstellung der Datensätze eine Mindestbioaktivität von 100µM (K_i- oder IC_{50}-Wert) verwendet. Zur besseren Vergleichbarkeit wurden die inSARa-Netzwerke jeweils mit den folgenden, gleichen Standardeinstellungen erzeugt: Mindest-MCS-Größe = 5 RG-Pseudoatome, Abbruch-Kriterium = 2% nicht-repräsentierte Moleküle, Ausschlussliste = aktiv.

Diese Analyse sollte der Untersuchung der typischen Verteilung von SAR-(Dis-)Kontinuität in inSARa-Netzwerken dienen. Zudem sollte hiermit untersucht werden, ob diese inSARa-basierte Charakterisierung mit der FP-basierten Datensatz-Charakterisierung korreliert. Zum Vergleich wurde daher für alle Datensätze der globale SAR-Index[135] mit Hilfe von SARANEA[136] berechnet (MACCS Keys, Tc-Schwellenwert: 0.75[135]) und die Korrelation auf Basis des Spearman-Rang-Korrelationskoeffizienten (vgl. Abschnitt 13.2) bestimmt.

17. Anwendung: Vergleich der inSARa-Netzwerke verschiedener Zielstrukturen

17.1. Zielsetzung: Ligandbasierte Analyse der Ähnlichkeit verschiedener Zielstrukturen mittels inSARa-Netzwerk-Vergleich

Abgrenzung zu bestehenden Ansätzen

Das ursprüngliche Ziel dieser Analyse war es, inSARa-Netzwerke verschiedener Zielstrukturen bezüglich Gemeinsamkeiten zu vergleichen. Im Kontext von Polypharmakologie und Chemogenomik-Methoden (vgl. Kapitel 8) lässt sich dieser Ansatz wie folgt einordnen:

Da inSARa-Netzwerke nur basierend auf Liganden-Information erstellt werden, handelt es hierbei um eine ligandbasierte Analyse. Das Besondere dieses im Vergleich zu anderen in der Literatur beschriebenen Ansätzen (vgl. Abschnitt 8.2) ist die Verwendung von (maximal) gemeinsamen Substrukturen als Vergleichsgrundlage der Ligandensätze. Im Gegensatz zu anderen Substruktur-Vergleichen[343–344] wird hierbei nicht die eigentliche Molekülstruktur verwendet, sondern aufgrund der RGs handelt es sich um einen Vergleich gemeinsamer pharmakophorer Eigenschaften verschiedener Targets. Bisher sind nur ligandbasierte Target-Analysen auf Basis pharmakophorer Fingerprints (z.B. CATS, FEPOPS, SHED)[346, 411] oder Target-Vorhersagen auf Basis von 3D-Pharmakophoren[360] beschrieben. Das nachfolgend vorgestellte Verfahren stellt somit einen neuartigen Ansatz im Bereich der Chemogenomik-Analyse dar.

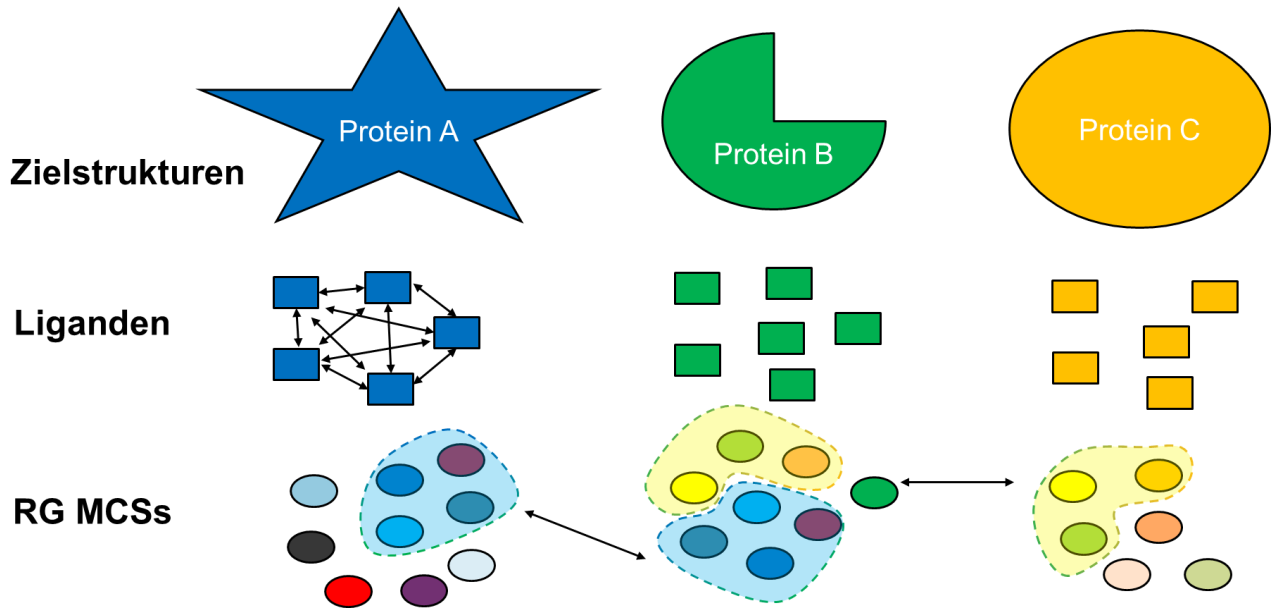


Abbildung 17.1. Prinzip der ligandbasierten Analyse der Ähnlichkeiten von Targets mittels RG-MCS-basierter inSARa-Netzwerke.

Prinzip des Ansatzes

Das dieser Analyse zugrunde liegende Prinzip wird in Abbildung 17.1 dargestellt. Der Vergleich der inSARa-Netzwerke erfolgt paarweise basierend auf den in den jeweiligen Netzwerken repräsentierten RG-MCSs. Diese resultieren ursprünglich aus dem paarweisen Vergleich aller Liganden, die zu dieser Zielstruktur gehören. Je größer der Anteil an gemeinsamen RG-MCSs zweier Targets ist, desto größer ist die Ähnlichkeit der Netzwerke. Auf diese Weise können wie bei anderen ligandbasierten Chemogenomik-Analysen indirekt Ähnlichkeiten zwischen Zielstrukturen hergestellt werden (SPP-Schlussfolgerung, vgl. Kapitel 8.1: ähnliche Liganden → ähnliche Eigenschaften → Bindung an ähnliche Zielstrukturen). In Abbildung 17.1 ist zu sehen, dass sich die RG-MCS-Mengen von Protein A und B deutlich überlappen, ebenso wie die MCS-Mengen von Protein B und C. Folglich besteht zwischen Protein A und B bzw. Protein B und C eine gewisse Ähnlichkeitsbeziehung.

Anwendungsmöglichkeiten

Solche Analysen von Target-Ähnlichkeiten können beispielsweise wertvolle Informationen über potentielle Kreuzreaktivitäten zwischen den entsprechenden Zielstrukturen liefern und der Vorhersage möglicher unerwünschten Wirkungen dienen. Von besonderem Interesse ist dies für Zielstrukturen, für die z.B. keine strukturelle Information vorhanden ist, jedoch aber eine Menge an aktiven Liganden bekannt ist. Da sich membranständige Proteine (u.a. GPCRs[412] oder Ionenkanäle[413]) nur schwer auskristallisieren lassen, fehlt für diese Zielstrukturen i.d.R. strukturelle Information. In Anbetracht der Tatsache, dass die On-Targets sogenannter „Blockbuster“ meist GPCRs sind bzw. mehr als 50% aller zugelassenen Arzneistoffe an GPCRs binden[412], liefern solche Analysen wertvolle Informationen bezüglich potentieller Off-Target-Beziehungen. Dies kann zur Vorhersage von potentiellen UAWs oder aber nach dem SOSA-Prinzip[313] (vgl. Abschnitt 8.1) zum „Drug Repurposing“ genutzt werden. Zum anderen können durch die Identifizierung ähnlicher Targets ggf. auch Ideen gewonnen werden, die für die Entwicklung neuer Wirkstoffe wichtig sein könnten (z.B. bezüglich selektivitätsentscheidener pharmakophorer Merkmale[414] oder potentiell neuer Liganden[246]).

17.2. Methode

Der Workflow der gesamten Analyse, die nachfolgend im Detail beschrieben wird, wird in Abbildung 17.3 veranschaulicht.

Schritt 1: Erzeugung eines inSARa-Netzwerkes für jedes Target

Im ersten Schritt wurde ausgehend von den vorbereiteten Datensätze, die jeweils Liganden mit einer bestimmten Mindest-Bioaktivität an der jeweiligen Zielstruktur enthalten, zunächst für jede Zielstruktur ein inSARa-Netzwerk wie in Kapitel 10 beschrieben mit den folgenden Standard-Einstellungen generiert: Mindest-MCS-Größe = 3, Stopp-Kriterium = 2% nicht-repräsentierte Moleküle. Zum Ausschluss von Ähnlichkeitsbeziehungen, die auf unspezifischer Zufalls-Ähnlichkeit beruhen, wurde die Ausschlussliste aktiv gesetzt (vgl. Kapitel 18.1).

Schritt 2: Bestimmung der gemeinsamen RG-MCSs in den inSARa-Netzwerken eines Target-Paares A und B unter Berücksichtigung von Substruktur-MCSs

Im nächsten Schritt wurden die inSARa-Netzwerke aller Targets auf Basis der in den Netzwerken auftretenden RG-MCSs, die die gemeinsamen pharmakophoren Eigenschaften der Liganden repräsentieren, verglichen. Dazu wurden paarweise die gemeinsamen RG-MCSs in den inSARa-Netzwerken der Zielstrukturen bestimmt. Beim Vergleich der Netzwerke sollte berücksichtigt werden, dass inSARa-Netzwerke verschiedener Zielstrukturen in der Regel eine unterschiedliche Anzahl an RG-MCSs bedingt durch Unterschiede in der Datensatzgröße und der Diversität der Liganden aufweisen. Da dieser Größenunterschied zu einer gewissen Verzerrung bei der nachfolgenden Berechnung der Ähnlichkeit führt, wurden zum Ausgleich zusätzlich noch *Substruktur-Beziehungen* bei der Bestimmung gemeinsamer MCSs *berücksichtigt* (vgl. Pseudocode in Abbildung 17.2).

Da ein MCS häufig die Superstruktur von weiteren kleineren MCSs in den inSARa-Netzwerken darstellt, wurde, sofern beide Targets einen gemeinsamen MCS aufweisen, bei *Berücksichtigung von Substruktur-Beziehungen* geprüft, ob dieser gemeinsame MCS die Superstruktur von weiteren MCSs in den beiden inSARa-Netzwerken bzw. MCS-Mengen beider Targets darstellt (siehe *mit (A) markierter Abschnitt in Abbildung 17.2*). Denn wenn der größere MCSs eine gemeinsame Substruktur von Liganden beider Targets darstellt, so gilt auch, dass kleinere MCSs ebenfalls eine gemeinsame Substruktur darstellen, auch wenn sie unter Umständen nicht in beiden MCS-Mengen bzw. inSARa-Netzwerken explizit auftauchen. Diese Substruktur-MCSs wurden für die Auswertung ebenfalls der Menge an gemeinsamen MCSs hinzugefügt (siehe mit (A) markierter Abschnitt in Abbildung 17.2). Umfangreiche Voranalysen haben gezeigt, dass sich hierdurch bedeutende deutlicher von unbedeutenden Ähnlichkeitsbeziehungen differenzieren lassen.

Soll die Ähnlichkeit zwischen zwei Targets *ohne Berücksichtigung von Substruktur-Beziehungen* berechnet werden, so wird nur geprüft, ob ein MCS aus der MCS-Menge von Target A auch in der MCS-Menge von Target B vorhanden ist. Ist dies der Fall so wird dieser

MCS der Menge gemeinsamer MCSs hinzugefügt so wird der *in Abbildung 17.2 mit (A) markierte Abschnitt*, der die Suchen nach Substruktur-MCSs repräsentiert, *übersprungen*.

Algorithm: Determine MCSs common to target A and B

Input:

MCS list of target A = { MCS_{a1}, MCS_{a2}, MCS_{a3}, …, MCS_{an} }

MCS list of target B = { MCS_{b1}, MCS_{b2}, MCS_{b3}, …, MCS_{bm} }

Output: MCS list only target A, MCS list only target B, common MCSs list of target AB

```
common MCSs list of targetAB ← ∅;
foreach MCS_ax in MCS list of target A do
    if MCS_ax in MCS list of target B then
        common MCSs list of target AB ← MCS_ax;
        remove MCS_ax from MCS list of target A;
        remove MCS_ax from MCS list of target B;
        foreach MCS_ay in MCS list of target A do
            if MCS_ay is substructure of MCS_ax then
                common MCSs list of targetAB ← MCS_ay;
                remove MCS_ay from MCS list of target A;
        foreach MCS_bz in MCS list of target B do
            if MCS_bz is substructure of MCS_ax then
                common MCSs list of targetAB ← MCS_bz;
                remove MCS_bz from MCS list of target B;
MCS list only target B ← MCS list of target A;
MCS list only target A ← MCS list of target B;
return MCS list only target A, MCS list only target B, common MCSs list target AB;
```

(A)

Abbildung 17.2. Pseudocode (in Englisch) für den Algorithmus zur Bestimmung der gemeinsamen MCSs von Target A und B unter Berücksichtigung von Substruktur-Beziehungen. Wird der mit (A) markierte Abschnitt übersprungen, so werden die gemeinsamen MCSs von Target A und B ohne Berücksichtigung von Substruktur-Beziehungen bestimmt.

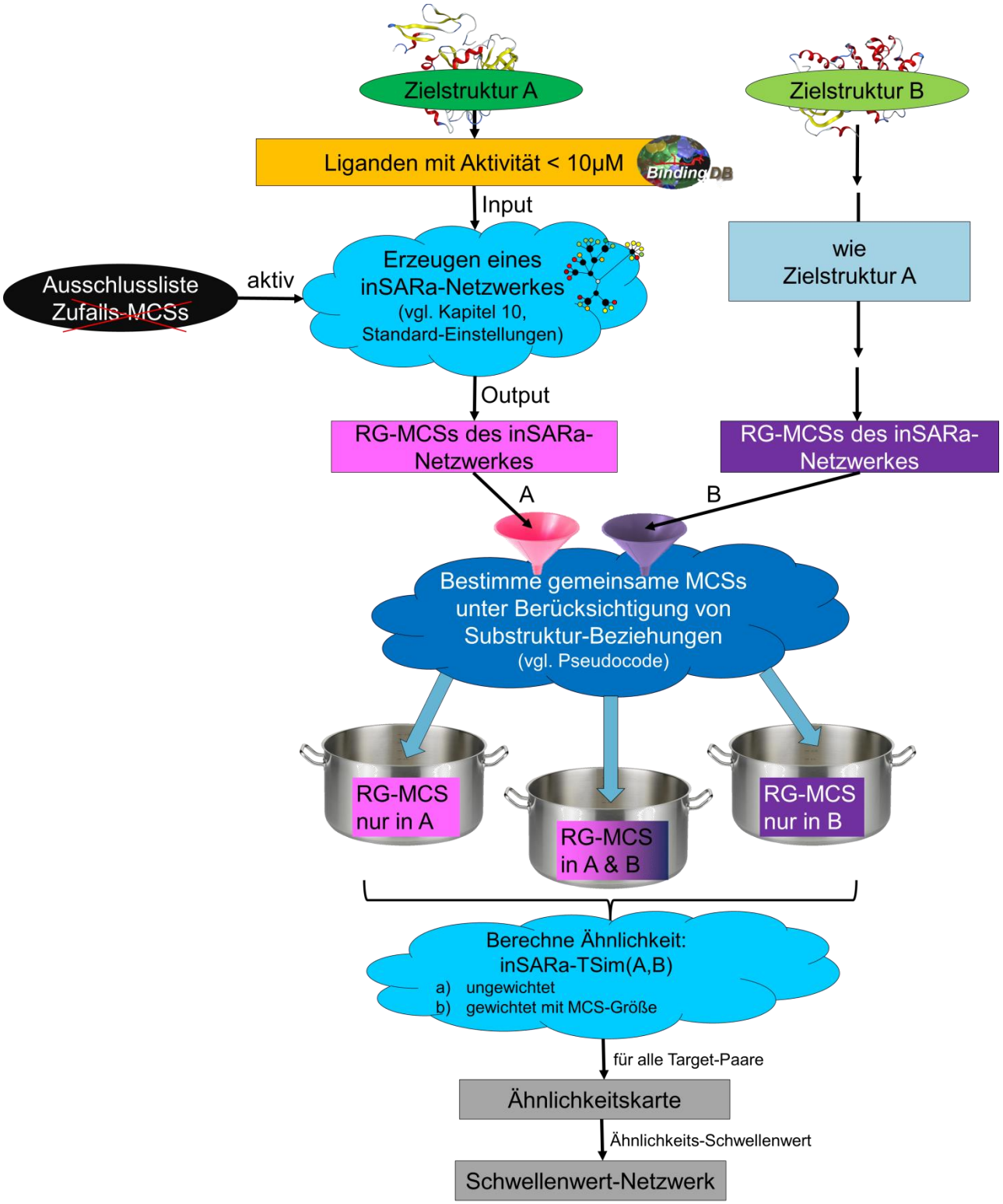


Abbildung 17.3. Workflow der Ligand-basierten Analyse von Target-Ähnlichkeiten unter Verwendung von inSARa-Netzwerken. Die Ähnlichkeit wird paarweise zwischen den Targets berechnet anhand der gemeinsamen Menge an RG-MCSs in den beiden inSARa-Netzwerken (weitere Details siehe Text).

Schritt 3: Paarweise Berechnung der Ähnlichkeit der inSARa-Netzwerke basierend auf den gemeinsamen RG-MCSs

Basierend auf der im vorangegangen Schritt bestimmten Menge an gemeinsamen RG-MCSs wurde für jedes Target-Paar A und B die Ähnlichkeit einerseits ungewichtet und andererseits gewichtet mit der MCS-Größe berechnet. Hierfür wurde ein vom Tanimoto-Koeffizienten (vgl. Abschnitt 2.3.3) abgeleiteter Ähnlichkeits-Wert inSARa-TSim nach der nachfolgenden Formel berechnet:

$$\text{inSARa-TSim(A,B)} = \frac{\sum_{i=1}^{c} \text{SizeAB}_i}{\sum_{i=1}^{c} \text{SizeAB}_i + \sum_{j=1}^{a} \text{SizeA}_j + \sum_{k=1}^{b} \text{SizeB}_k} \qquad (17.1)$$

a bzw. *b* gibt die Gesamtzahl an MCSs, die nur in dem inSARa-Netzwerk von Target A bzw. B vorkommen. *c* ist die Gesamtzahl an gemeinsamen MCSs von A und B. Für die ungewichtete Berechnung der Ähnlichkeit wird SizeA, SizeAB und SizeB jeweils gleich 1 gesetzt. Um zu berücksichtigen, dass größere MCSs eine größere Ähnlichkeit zwischen Liganden bedeuten, kann die MCS-Größe bei der Berechnung der Ähnlichkeit berücksichtigt werden. SizeAB stellt hierbei die RG-MCS-Größe des *i*-ten MCS dar, den die Targets A und B gemeinsam haben. SizeA bzw. SizeB ist die RG-MCS-Größe des *j*-ten bzw. *k*-ten MCS, der nur in dem inSARa-Netzwerk von Target A bzw. B vorkommen. So kann zum einen Rauschen reduziert werden, das sich aus der Gemeinsamkeit von sehr kleinen MCSs ergeben kann. Jedoch kann durch die Gewichtung auch bedeutsame Ähnlichkeit schwerer erkannt werden. Denn wie unter 18.3.1 gezeigt, weisen bestimmte Datensätze deutliche Unterschiede in der Größe der RGs auf. Durch die starke Gewichtung von großen MCSs sinkt die Ähnlichkeit zwischen Targets mit deutlichen Unterschieden in der RG-Größe. Im Folgenden wurde für die Analysen daher sowohl der gewichtete als auch der ungewichtete inSARa-TSim berücksichtigt. Der Wert des inSARa-TSim wird anschließend durch Multiplikation mit 100 auf den Bereich von 0 bis 100 skaliert.

Untersucht wurde auch die Verwendung der Formel des kontinuierlichen Tanimotos (vgl. Formel 2.4 in Abschnitt 2.3.3). Hierfür würde bei der obigen Formel die einzelnen Summanden noch quadriert werden (nachfolgend daher als „quadrierter Tc“ bezeichnet). Die Analysen haben ergeben, dass eine Korrelation von 1.0 (Verwendung des Spearman-Rang-Korrelationskoeffizienten ρ_S (vgl. Abschnitt 13.2), da kein linearer Zusammenhang und keine Normalverteilung der Daten angenommen werden kann) der Ähnlichkeitswerte des gewichteten inSARa-TSim und des quadrierten Tc (skaliert jeweils auf den Bereich von 0 bis 100) für die analysierten Targets besteht. Zudem werden im Allgemeinen geringere Ähnlichkeitswerte für den quadrierten Tc bestimmt (vgl. Abbildung 26.2 im Anhang). Dies erschwert eine Unterscheidung von unbedeutendem Rauschen und bedeutsamer Ähnlichkeit. Aus diesen Gründen wurde für die nachfolgend gezeigten Analysen nur der gewichtete inSARa-TSim verwendet. Eine Verwendung des quadrierten Tc als Ähnlichkeits-Koeffizienten ist jedoch möglich. Eine Anpassung des Schwellenwertes ist in diesem Fall jedoch empfehlenswert.

17.3. Visualisierung

a) Ähnlichkeitskarte

Zur Visualisierung wurden die inSARa-TSim-Ähnlichkeitswerte der paarweisen Target-Vergleiche in Form einer Heatmap dargestellt. In dieser Ähnlichkeitskarte wird zur Visualisierung der Stärke der Ähnlichkeit ein Farbschema von Weiß über Blau und Grün nach Rot verwendet, wobei weiße Stellen in der Karte anzeigen, dass keine oder nur sehr geringfügige Überlappung in der RG-MCSs-Menge der beiden Targets besteht. Blau eingefärbte Kästen deuten stärkere Überlappung an. Grün zeigt sehr starke und Rot maximale Ähnlichkeit (vgl. Eigenvergleich auf Diagonale) an. Bei der Ähnlichkeitskarte ist zu beachten, dass sie zur Hälfte redundante Information enthält.

b) Schwellenwert-Netzwerk

Da die in der Heatmap dargestellten Ähnlichkeitsbeziehungen aufgrund der großen Datenzahl nur schwer auf einen Blick zu erfassen ist, wurde zusätzlich ein Schwellenwert-Netzwerk erstellt. Diese Art von Netzwerk ist analog zu den NSGs (vgl. Kapitel 2.6) und folgendermaßen erzeugt:

1.) Jedes Target wird durch einen einzelnen Knoten repräsentiert.

2.) Eine Kante wird dann zwischen zwei Knoten erzeugt, sofern der berechnete paarweise Ähnlichkeits-Wert dieser Targets größer oder genauso groß wie ein bestimmter, festgelegter Ähnlichkeits-Schwellenwert ist.

3.) Jedes Target wird einer der folgenden Target-Klassen zugeordnet: Protease, Kinase, andere Enzyme, GPCR, nukleärer Rezeptor, Transporter, Ionenkanal, Orphan. Jeder Knoten wird dann entsprechend der Target-Klasse, zu der das repräsentierte Target gehört, eingefärbt.

Diese Art der Darstellung ermöglicht das leichtere Erkennen von (un)erwartete Ähnlichkeits-Beziehungen in der großen Menge analysierter Zielstrukturen.

Als Schwellenwert für Netzwerke, die auf dem gewichteten inSARa-TSim-Ähnlichkeits-Wert basieren, wurde empirisch ein Ähnlichkeitswert von 3.0 festgelegt. Es ist anzunehmen, dass in der Mehrheit der paarweisen Vergleiche immer geringfügige Ähnlichkeiten gefunden werden. Hierbei kann unbedeutendes „Untergrundrauschen“ angenommen werden. Eine Analyse hat gezeigt, dass der Median aller Werte bei 0.22 (Mean: 0.93) liegt Ein Schwellenwert von 3.0 hebt sich somit deutlich davon ab. Zudem hat sich gezeigt, dass bei diesem Wert zu erwartende (d.h. sinnvolle) Ähnlichkeits-Beziehungen resultieren.

17.4. Analyse und Auswertung

1.) Untersuchung des Einflusses von Größenunterschieden und Gewichtung

Da für verschiedene Zielstrukturen eine unterschiedliche Zahl an Liganden bekannt bzw. veröffentlicht sind, weisen ligandbasierte Analysen gewisse Verzerrungen auf (vgl. auch KEISER et al.[346]). Bei dieser Analyse sind Unterschiede in den MCSs-Mengen der verschiedenen Targets unvermeidbar. Selbst bei Datensätzen, die gleich viele Liganden enthalten, resultiert aufgrund unterschiedlicher Diversität i.d.R. eine unterschiedliche Anzahl an MCSs mit der zusätzlichen Schwierigkeit der unterschiedlichen MCS-Größenverteilung.

Zur Untersuchung des Einflusses dieser Unterschiede auf den inSARa-TSim-Ähnlichkeits-Wert dienten die nachfolgend beschriebenen zwei Experimente. Als Datengrundlage für diese Experimente wurden repräsentativ drei bereits bekannte Datensätze (vgl. Abschnitt 11.2, THR, P38 und COX2) mit großen MCS-Mengen (> 1000 MCSs) ausgewählt.

a) „Selbst-Ähnlichkeits-Test“

Um den Einfluss des Größenunterschiedes im MCS-Pool zweier Zielstrukturen zu analysieren, wurden aus jedem der drei Gesamt-Datensätze zufällig unterschiedlich große Stichproben gezogen (vgl. Tabelle 17.1). Eine einzelne Stichprobe liefert jedoch eine wenig zuverlässige Aussage (ggf. Ausreißer). Um Informationen über den Streuungsbereich zu erhalten, wurde die Stichprobenziehung daher 10-mal wiederholt. Die unterschiedliche MCS-Größen-Verteilung wurde bei der Stichprobenziehung nicht berücksichtigt.

Für jede Zielstruktur wurden der Gesamtdatensatz und die verschiedenen Stichproben paarweise miteinander verglichen und die Ähnlichkeit wie unter 17.2 beschrieben berechnet. Dieser „Selbst-Ähnlichkeits-Test“ hat den Vorteil, dass eine existierende Ähnlichkeitsbeziehung zwischen den verschieden großen untersuchten Datensatz-Paaren sichergestellt ist. Es gibt einen Hinweis auf den systematischen Fehler, der aus dem Vergleich von großen und kleinen MCS-Mengen resultiert. Des Weiteren sollte hiermit auch der Einfluss der Berücksichtigung von Substruktur-Beziehungen (Vergleich mit und ohne Einbeziehung von Substruktur-MCSs) analysiert werden.

Tabelle 17.1. Übersicht über Stichprobenziehung für „Selbst-Ähnlichkeits-Test“

Target	**THR**	**P38**	**COX2**
Gesamt-MCS-Menge	2732	1650	1132
Größe der zufällig gezogenen Stichproben (10 Wiederholungen)	2500, 2000, 1500, 1000, 750, 500, 250, 150, 100, 50	1500, 1250, 1000, 750, 500, 250, 150, 100, 50	1000, 750, 500, 250, 150, 100, 50

b) „Wiederfindungs-Test“

Beim „Selbst-Ähnlichkeits-Test“ stellt die gesamte Stichprobe eine Untermenge der Gesamtmenge dar. In normalen Datensätzen ist diese Bedingung normalerweise nicht erfüllt. Stattdessen stellt nur ein geringer Teil des einen Datensatzes eine Untermenge des anderen Datensatzes dar. Um dies zu simulieren, wurden die kleinsten Stichproben (50 MCSs) aus dem unter a) beschriebenen Versuchsaufbau jeweils genommen und dem Gesamt-Pool der jeweils anderen beiden Zielstrukturen hinzugefügt. So soll untersucht werden, ob auch kleine Ähnlichkeiten in sehr großen Datensätzen gefunden werden können.

2.) Beurteilung von Ähnlichkeits-Beziehungen

Die resultierenden Ähnlichkeits-Beziehungen wurden auf Basis der recherchierten Target-Klassen/Sub-Klassen analysiert (vgl. Tabelle 17.2).

In Anlehnung an Analysen anderer Gruppen[343, 349, 344] wurde zusätzlich ein Vergleich mit Sequenz-basierter, phylogenischer Ähnlichkeit durchgeführt. Da ein Targetklassen-übergreifender Vergleich wenig sinnvoll ist und Analysen selbst innerhalb einer Klasse aufgrund der in Abschnitt 8.2 genannten Gründe z.T. mit Schwierigkeiten verbunden ist, wurden hierfür keine eigenen BLAST-Analysen durchgeführt, sondern auf publizierte Sequenz-Vergleiche innerhalb der Klasse der GPCRs[343, 349, 415], Kinasen[336, 416], Proteasen (MEROPS[337–338, 417]), und Nukleären Rezeptoren[418] zurückgegriffen. Für Enzyme wurde zudem die recherchierte EC-Klassifizierung[419], die eine Einteilung nicht basierend auf Ähnlichkeit, sondern nach Funktion darstellt, als Vergleichsgrundlage verwendet.

Auf einen Vergleich mit strukturbasierter Ähnlichkeit wurde aufgrund der Schwierigkeit der korrekten Definition der Bindetasche und der eingeschränkten Verfügbarkeit von Kristallstruktur-Information für die analysierten Targets verzichtet.

Aufgrund der unterschiedlichen Datengrundlage ist ein direkter Vergleich mit Analysen anderer Gruppen schwierig. Daher wurden die sich ergebenden Target-Ähnlichkeits-Beziehungen im RG-MCS-basierten Netzwerk in Bezug auf vergleichbare oder gegensätzliche Trends mit den Ergebnissen bzw. Target-Netzwerken von PAOLINI et al.[342], KEISER et al.[345] und SUTHERLAND et al.[344] verglichen. Diese wurden ausgewählt, da es sich hierbei ebenfalls um ligandbasierte Analysen handelt, jedoch unterschiedliche Verfahren des Ähnlichkeitsvergleichs angewendet werden (direkter Vergleich gleicher Liganden[342], Fingerprint-basierte Ähnlichkeit[345] und Substruktur-basierte Ähnlichkeit[344], vgl. Kapitel 8.2).

3.) Validierung von Kreuzreaktivitäten

Alternativ zur sehr aufwändigen experimentellen Validierung wurde zur Beurteilung unerwarteter Verknüpfungen zwischen nicht-verwandten Targets zum einen mittels Literatur-Recherche als auch durch Suche in Bioaktivitätsdatenbanken (BindingDB und ChEMBL) nach beschriebenen Testungen von Liganden (insbesondere bekannter Arzneistoffe) an beiden Targets gesucht.

4.) Vergleich: Verwendung der gesamten MCS-Menge

Die in inSARa-Netzwerken repräsentierte MCS-Menge stellt nur einen bestimmten Anteil der gesamten MCS-Menge eines Targets dar. Daher stellt sich die Frage, ob und in welchem Umfang sich die Ergebnisse dieser Analyse unterscheiden, wenn die gesamte MCS-Menge eines Targets als Datengrundlage verwendet wird.

Die gesamte MCS-Menge wird gebildet von allen einzigartigen MCSs, die sich aus dem paarweisen Vergleich aller Moleküle eines Datensatzes (bzw. deren RGs) ergeben (vgl. Abschnitt 10.3). Die MCS-Menge der inSARa-Netzwerke wird hingegen gebildet von dem Anteil an MCSs aus dieser Gesamt-Menge, der benötigt wird, um einen vorgegebenen Anteil an Datensatz-Molekülen im inSARa-Netzwerk zu repräsentieren (vgl. Abschnitt 10.4). Im inSARa-Netzwerk sind daher nur diejenigen, ausgewählten Wurzel-MCSs und die jeweiligen zugehörigen Superstruktur-MCSs (d.h. MCSs, die eine Superstruktur des Wurzel-MCS darstellen) zu finden, die für eine ausreichende Abdeckung des chemischen Raumes eines Targets (definiert über den Anteil nicht-repräsentierter Datensatz-Moleküle) notwendig sind. Die inSARa-MCS-Menge ist v.a. um kleinere MCSs, die zumeist Wurzel-MCSs darstellen, im Vergleich zur gesamten MCS-Menge reduziert.

Zur Untersuchung des Einflusses des oben beschriebenen Unterschiedes in der MCS-Menge wurde die Analyse auf Basis der gesamten MCS-Menge der jeweiligen Datensätze unter deutlich erhöhtem Rechenaufwand wiederholt. Zum Vergleich wurde anschließend die Korrelation der Ähnlichkeitswerte auf Basis der inSARa-MCSs und der gesamten MCS-Menge (aufgrund der fehlenden Normalverteilung) unter Verwendung des Spearman-Rang-Korrelationskoeffizienten bestimmt (vgl. auch HERT et al.[346]). Zudem wurde die prozentuale Übereinstimmung des Schwellenwert-Netzwerkes bei verschiedenen Schwellenwerten im Vergleich zum inSARa-basierten Schwellenwert-Netzwerk untersucht (berechnet als % gleiche Verknüpfungen bei gleicher Knotenzahl, vgl. HERT et al.[346]).

17.5. Datengrundlage

Für die Analyse wurden nur Datensätze aus der BindingDB verwendet. Die Datensatz-Auswahl erfolgte nach den folgenden Kriterien:

1.) Analog zu Abschnitt 11.2 wurden nur Datensätze, für die K_i- oder IC_{50}-Werte als Bioaktivitätswerte verfügbar sind, berücksichtigt.

2.) In Anlehnung an Chemogenomik-Analysen anderer Gruppen[342–343, 349–350] wurde im Gegensatz zu den SAR-Analysen einzelner Targets eine Mindest-Bioaktivität < 10 µM für die Datensatz-Zusammenstellung verwendet. Molekülen mit K_i- oder IC_{50}-Werte größer oder gleich 10 µM wurden in der Analyse nicht berücksichtigt. Dieser Wert ist auch dadurch zu begründen, dass Pharmafirmen bei ihren Sicherheits-Testungen potentieller Arzneistoffkandidaten ein Molekül mit einem IC_{50}-Wert < 10 µM oftmals als Off-Target-„Hit“ definieren[310].

3.) Um möglichst viel Liganden-Information zu haben, wurden große Datensätze bevorzugt ausgewählt. Die ausgewählten Roh-Datensätze sollten aus mindestens 500 Molekülen bestehen (geringe Anzahl an Ausnahmen, vgl. * in Tabelle 17.2). Da bei der Aufbereitung der Roh-Datensätze noch einige Moleküle (u.a. aufgrund bestimmter Eigenschaften, Duplikate oder zu geringe Bioaktivität) ausgeschlossen werden, schrumpfen die bereinigten Datensätze im Vergleich zu den Roh-Datensätzen deutlich (vgl. Tabelle 17.2).

4.) Es wurden möglichst diverse Targets ausgewählt, die die verschiedenen Protein-Familien (Enzyme, GPCRs, nukleäre Rezeptoren, Liganden-gesteuerte Ionenkanäle, Transporter), Klassen (bei den Enzymen z.B. Proteasen, Kinasen, Oxidoreduktase etc.) und Subklassen (bei den Proteasen z.B. Serinproteasen, Aspartatproteasen, Cysteinproteasen, Metalloproteasen) repräsentieren.

5.) Bekannte pharmakologische Targets oder Targets mit gut untersuchten SARs oder bekannten Pharmakophoren wurden bevorzugt ausgewählt, da hier am meisten Literatur zur Validierung von hergestellten Ähnlichkeits-Beziehungen zu erwarten ist.

6.) Ziel war es mindestens 100 Targets, die diese Kriterien erfüllen, zu sammeln, um eine möglichst große Stichprobe für die Analyse zu haben, in der die wichtigsten Targets vertreten sind.

Anhand dieser Kriterien wurden 140 Datensätze manuell aus der BindingDB ausgewählt und heruntergeladen. Details zu den Datensätzen sind in der Tabelle 17.2 zu finden. Anschließend wurden die einzelnen Datensätze wie unter Kapitel 12 beschrieben vorbereitet.

Tabelle 17.2. Übersicht über verwendete Zielstrukturen/Datensätze.

Nr.	Abkürzung	Target Name	Protein-Familie	Klasse	Sub-Klasse (MEROPS Familie/Clan) (EC-Code)	Bioaktivitäts-Wert	Anzahl an Datensatz-Molekülen nach Download	Anzahl an Molekülen nach Filtern und RG-Generierung	Gesamtmenge an MCSs	Menge an inSARa-MCSs
(1)	FXa	Koagulations-Faktor Xa	Enzym	Protease	Serin-Protease (S1A/PA)	IC_{50}	2064	1637	1686	973
(2)	DPP-II	Dipeptidyl-peptidase-II	Enzym	Protease	(Serine-)Protease (S28/SC)	IC_{50}	1028	325	235	155
(3)	DPP-IV	Dipeptidyl-peptidase-IV	Enzym	Protease	(Serine-)Protease (S9B/SC)	K_i	582	452	260	181
(4)	Elastase-2	Elastase-2 (Leukozyten-Elastase)	Enzym	Protease	(Serine-)Protease (S1A/PA)	K_i	925	635	509	381
(5)	PEP	Prolylendo-peptidase/ Prolyloligo-peptidase	Enzym	Protease	Serin-Protease (S9A/SC)	IC_{50}	626	326	173	138
(6)	Plasmin	Plasmin	Enzym	Protease	Serin-Protease (S1A/PA)	K_i	832	317	337	207
(7)	tPA	tissue-type plasminogen activator	Enzym	Protease	Serin-Protease (S1A/PA)	K_i	412	110	115	73
(8)	Thrombin	Thrombin	Enzym	Protease	Serin-Protease (S1A/PA)	K_i	6844	2570	3640	2732
(9)	Trypsin	Trypsin	Enzym	Protease	Serin-Protease (S1A/PA)	K_i	1659	869	1198	799
(10)	uPA	Urokinase-Typ Plasminogen Aktivator	Enzym	Protease	Serin-Protease (S1A/PA)	K_i	1217	457	406	276
(11)	BACE-1	ß-Sekretase (Memapsin 1)	Enzym	Protease	Aspartat-Protease (A1/AA)	IC_{50}	3913	1845	2788	1934

(12)	CatD	Cathepsin D	Enzym	Protease	Aspartat-Protease (A1/AA)	IC_{50}	1248	670	1037	691
(13)	HIV-1-Protease	HIV-1 Protease (HIV-1 retropepsin)	Enzym	Protease	Aspartat-Protease (A2/AA)	IC_{50}	6112	2605	2897	2300
(14)	Renin	Renin	Enzym	Protease	Aspartat-Protease (A1/AA)	IC_{50}	3635	1605	2855	2315
(15)	CatB	Cathepsin B	Enzym	Protease	Cystein-Protease (C1/CA)	IC_{50}	1677	721	462	348
(16)	CatK	Cathepsin K	Enzym	Protease	Cystein-Protease (C1/CA)	IC_{50}	1503	1073	780	518
(17)	CatL	Cathepsin L	Enzym	Protease	Cystein-Protease (C1/CA)	IC_{50}	2238	713	741	507
(18)	CatS	Cathepsin S	Enzym	Protease	Cystein-Protease (C1/CA)	IC_{50}	1697	1076	1118	758
(19)	Cruzipain	Cruzipain	Enzym	Protease	Cystein-Protease (C1/CA)	IC_{50}	578	178	158	101
(20)	Caspase-1	Caspase-1	Enzym	Protease	Cystein-Protease (C14/CD)	IC_{50}	605	338	384	299
(21)	Caspase-3	Caspase-3	Enzym	Protease	Cystein-Protease (C14/CD)	IC_{50}	1677	721	757	542
(22)	ACE	Angiotensin-converting enzyme (Peptidyl-dipeptidase A)	Enzym	Protease	Metalloprotease (M2/MA)	IC_{50}	698	438	345	224
(23)	MMP-1	Matrix-metallo-peptidase-1	Enzym	Protease	Metalloprotease (M10/MA)	IC_{50}	2450	1404	1525	1016
(24)	MMP-13	Matrix-metallo-peptidase-13	Enzym	Protease	Metalloprotease (M10/MA)	IC_{50}	2186	1699	1588	1010
(25)	ACAT-1	Acetyl-CoA acetyltransferase	Enzym	andere Enzyme	(Acyl-)Transferase (EC=2.3.1.9)	IC_{50}	1076	719	391	259
(26)	AChE	Acetylcholinesterase	Enzym	andere Enzyme	Hydrolase (Esterase)	K_i	625	194	216	156

					(EC=3.1.1.7)					
(27)	Aromatase	Aromatase (Cytochrome P450 19A1, Testosteron-Monooxygenase, Estrogen-Synthase)	Enzym	andere Enzyme	Oxidoreduktase (EC=1.14.14.14)	IC_{50}	1917	1096	750	558
(28)	BChE	Butyrylcholin-esterase (Pseudocholin-esterase)	Enzym	andere Enzyme	Hydrolase (Esterase) (EC=3.1.1.8)	IC_{50}	2118	1004	803	555
(29)	COX-1	Cyclooxygenase-1 (Prostaglandin-G/H-Synthase)	Enzym	andere Enzyme	Oxidoreduktase (EC=1.14.99.1)	IC_{50}	4115	932	912	671
(30)	COX-2	Cyclooxy-genase-2 (Prostaglandin-G/H-Synthase)	Enzym	andere Enzyme	Oxidoreduktase (EC=1.14.99.1)	IC_{50}	4357	2020	1452	1132
(31)	CA-1	Carbo-anhydrase-1	Enzym	andere Enzyme	Lyase (EC=4.2.1.1)	K_i	3654	1751	1489	1152
(32)	CA-2	Carbo-anhydrase-2	Enzym	andere Enzyme	Lyase (EC=4.2.1.1)	K_i	3949	2105	1714	1356
(33)	DHFR	Dihydrofolat-Reduktase	Enzym	andere Enzyme	Oxidoreduktase (EC=1.5.1.3)	IC_{50}	4133	663	643	334
(34)	HMG-CoA-Red	HMG-CoA-Reduktase (3-hydroxy-3-methylglutaryl-coenzyme A reductase)	Enzym	andere Enzyme	Oxidoreduktase (EC=1.1.1.34)	IC_{50}	986	667	532	255
(35)	HIV-1-Integrase	HIV Typ-1 Integrase	Enzym	Andere Enzyme	(Nucleotidyl-) Transferase (EC=2.7.7.?)	IC_{50}	2358	581	644	403

(36)	HIV-1-RT	HIV Typ-1 Reverse Transkriptase	Enzym	Andere Enzyme	(Nucleotidyl-) Transferase (EC=2.7.7.49)	IC_{50}	1531	1115	676	506
(37)	5-LOX	Arachidonate 5-lipoxygenase	Enzym	andere Enzyme	Oxidoreduktase (EC=1.13.11.34)	IC_{50}	2220	1291	1122	822
(38)	MAO-A	Monoamino-oxidase Typ A	Enzym	andere Enzyme	Oxidoreduktase (EC=1.4.3.4)	IC_{50}	1852	861	562	485
(39)	MAO-B	Monoamino-oxidase Typ B	Enzym	andere Enzyme	Oxidoreduktase (EC=1.4.3.4)	IC_{50}	1232	744	390	314
(40)	PDE-3	Phosphodi-esterase-3	Enzym	Andere Enzyme	Hydrolase (Esterase) (EC=3.1.4.17)	IC_{50}	994	459	351	234
(41)	PDE-4	Phosphodi-esterase-4	Enzym	Andere Enzyme	Hydrolase (Esterase) (EC=3.1.4.17)	IC_{50}	1213	756	1002	706
(42)	PDE-5	Phosphodi-esterase-5	Enzym	Andere Enzyme	Hydrolase (Esterase) (EC=3.1.4.17)	IC_{50}	1303	994	1191	853
(43)	PDE-10a	Phosphodi-esterase-10a	Enzym	Andere Enzyme	Hydrolase (Esterase) (EC=3.1.4.17)	IC_{50}	895	826	531	371
(44)	PG-E-Synthase	Prostaglandin-E-Synthase	Enzym	andere Enzyme	Isomerase (EC=5.3.99.3)	IC_{50}	630	286	360	246
(45)	PL-A2	Phospholipase-A2	Enzym	andere Enzyme	Hydrolase (Esterase) (EC=3.1.1.4)	IC_{50}	588	291	269	216
(46)	PT-Phosphatase -1B	Protein-Tyrosin-Phosphase Typ 1B	Enzym	Andere Enzyme	Hydrolase (Esterase) (EC=EC 3.1.3.48) (Phosphorsäure-ester)	IC_{50}	2412	1027	1458	1040
(47)	TS	Thymidylat-Synthase	Enzym	andere Enzyme	(Methyl-) Transferase (EC=2.1.1.45)	IC_{50}	1012	271	172	115
(48)	XDH	Xanthin-Oxidase	Enzym	andere	Oxidoreduktase	IC_{50}	387	96	60	56

		(Xanthin-Dehydrogenase)		Enzyme	(EC=1.17.3.2)					
(49)	ABL-1	Tyrosine-protein kinase ABL-1	Enzym	Kinase	TK-Gruppe (EC=2.7.10.2) (Tyrosin-Kinasen)	IC_{50}	662	333	485	287
(50)	AKT-1	RAC-alpha serine/threonine-protein kinase (Protein kinase B)	Enzym	Kinase	AGC-Gruppe (EC=2.7.11.1) (enthält PKA, PKG, PKC-Familien)	IC_{50}	1876	894	771	472
(51)	Aurora-A	Aurora kinase-A	Enzym	Kinase	Andere (EC=2.7.11.1)	IC_{50}	1257	835	908	617
(52)	CDK1	Cyclin-dependent kinase-1	Enzym	Kinase	CMGC-Gruppe (EC=2.7.11.22) (enthält CDK, MAPK, GSK3, CLK Familien)	IC_{50}	1648	885	840	612
(53)	CDK2	Cyclin-dependent kinase-2	Enzym	Kinase	CMGC-Gruppe (EC=2.7.11.22)	IC_{50}	2557	1395	1326	887
(54)	CDK4	Cyclin-dependent kinase-4	Enzym	Kinase	CMGC-Gruppe (EC=2.7.11.22)	IC_{50}	1313	676	571	398
(55)	CDK5	Cyclin-dependent kinase-5	Enzym	Kinase	CMGC-Gruppe (EC=2.7.11.22)	IC_{50}	499	260	155	113
(56)	CHK-1	Serine/threonine-protein kinase Chk1 (Checkpoint kinase 1)	Enzym	Kinase	CAMK-Gruppe (EC=2.7.11.1) (Calcium/ calmodulin-dependent protein kinase)	IC_{50}	1879	1224	1124	791
(57)	EGFR	Epidermal growth factor receptor	Enzym	Kinase	TK-Gruppe (EC=2.7.10.1)	IC_{50}	4648	2303	2270	1611
(58)	FGFR-1	Fibroblast growth factor receptor 1	Enzym	Kinase	TK-Gruppe (EC=2.7.10.1)	IC_{50}	905	519	509	347
(59)	GSK-3	Glycogen synthase kinase-	Enzym	Kinase	CMGC-Gruppe (EC=2.7.11.26)	IC_{50}	1666	1077	1041	774

		3								
(60)	IGFR-1	Insulin-like growth factor receptor 1	Enzym	Kinase	TK-Gruppe (EC=2.7.10.1)	IC_{50}	750	427	369	243
(61)	Insulin-Rez	Insulin receptor	Enzym	Kinase	TK-Gruppe (EC=2.7.10.1)	IC_{50}	717	342	421	296
(62)	JNK-3	c-Jun N-terminal kinase 1 (Mitogen-activated protein kinase 10)	Enzym	Kinase	CMGC-Gruppe (EC= 2.7.11.24)	IC_{50}	1012	497	524	362
(63)	mTOR	Serine/threonine-protein kinase mTOR (Mammalian target of rapamycin)	Enzym	Kinase	Andere (PI3K-Familie) (EC=2.7.11.1)	IC_{50}	1301	831	866	637
(64)	P38	Mitogen-activated protein kinase p38 alpha (Mitogen-activated protein kinase 14)	Enzym	Kinase	CMGC-Gruppe (EC=2.7.11.24)	IC_{50}	3272	2357	2363	1650
(65)	PDGFR-beta	Platelet-derived growth factor receptor beta	Enzym	Kinase	TK-Gruppe (EC=2.7.10.1)	IC_{50}	1792	890	862	588
(66)	PI3K	Phosphatidylinositide 3-kinase	Enzym	Kinase	Andere (PI3K-Familie) (EC=2.7.11.X)	IC_{50}	1821	1296	1335	963
(67)	PLK-1	Serine/threonine-protein kinase PLK1 (Polo-like kinase 1)	Enzym	Kinase	Andere (EC=2.7.11.21)	IC_{50}	738	306	276	194
(68)	SRC	Proto-oncogene tyrosine-protein	Enzym	Kinase	TK-Gruppe (EC=2.7.10.2)	IC_{50}	2656	1106	1626	1146

		kinase Src								
(69)	VEGFR-2	Vascular endothelial growth factor receptor-2	Enzym	Kinase	TK-Gruppe (EC=2.7.10.1)	IC_{50}	4629	2613	2767	2067
(70)	Androgen	Androgen-Rezeptor	Nukleärer Rezeptor		3C-Subfamilie	IC_{50}	1460	716	518	391
(71)	Gluccocortic oid	Gluccocorticoid-Rezeptor	Nukleärer Rezeptor		3C-Subfamilie	IC_{50}	1368	686	667	506
(72)	Estrogen-alpha	Estrogen-Rezeptor alpha	Nukleärer Rezeptor		3A-Subfamilie	IC_{50}	2109	1090	863	605
(73)	Estrogen-beta	Estrogen-Rezeptor beta	Nukleärer Rezeptor		3A-Subfamilie	IC_{50}	2150	1231	1112	749
(74)	LXR	Liver-X-Rezeptor	Nukleärer Rezeptor		1H-Subfamilie	IC_{50}	628	398	312	218
(75)	Mineralocorti coid	Mineralocorticoid -Rezeptor	Nukleärer Rezeptor		3C-Subfamilie	IC_{50}	477	325	300	228
(76)	Progesteron	Progesteron-Rezeptor	Nukleärer Rezeptor		3C-Subfamilie	IC_{50}	1366	969	833	619
(77)	PPAR-gamma	Peroxisome proliferator-activated Rezeptor gamma	Nukleärer Rezeptor		1C-Subfamilie	IC_{50}	1280	827	788	551
(78)	TR-beta	Thyroid-Hormon Rezeptor beta	Nukleärer Rezeptor		1A-Subfamilie	IC_{50}	622	305	141	112
(79)	Adenosin-A1	Adenosin A1 Rezeptor	GPCR	Adenosin		K_i	2554	1659	1288	927
(80)	Adenosin-A2A	Adenosin A2a Rezeptor	GPCR	Adenosin		K_i	2049	1349	978	639
(81)	Adenosin-A3	Adenosin A3 Rezeptor	GPCR	Adenosin		K_i	4082	2356	2087	1467
(82)	Alpha-1A	Alpha-1a adrenerger Rezeptor	GPCR	aminerg		K_i	1956 871	1327 485	946 665	447 317
(83)	Alpha-2C	Alpha-2c	GPCR	aminerg		K_i	623	366	348	238

(84)	Beta-1	Beta-1 adrenerger Rezeptor	GPCR	aminerg		IC_{50}	584 612	417 420	265 268	182 182
(85)	Beta-2	Beta-2 adrenerger Rezeptor	GPCR	aminerg		K_i	327	99	132	98
(86)	Beta-3	Beta-3 adrenerger Rezeptor	GPCR	aminerg		K_i	310	255	174	135
(87)	D1	Dopamine D1 Rezeptor	GPCR	aminerg		K_i	2611	1076	1052	724
(88)	D2	Dopamine D2 Rezeptor	GPCR	aminerg		K_i	3330	2114	1575	1171
(89)	D3	Dopamine D3 Rezeptor	GPCR	aminerg		K_i	2932	1983	1425	1090
(90)	D4	Dopamine D4 Rezeptor	GPCR	aminerg		K_i	2227	1504	1027	751
(91)	H1	Histamine H1 Rezeptor	GPCR	aminerg		K_i	1168	532	489	352
(92)	H3	Histamine H3 Rezeptor	GPCR	aminerg		K_i	3753	2337	1356	875
(93)	H4	Histamine H3 Rezeptor	GPCR	aminerg		K_i	788	481	244	175
(94)	5-HT-1a	Serotonin 1a Rezeptor	GPCR	aminerg		K_i	2950	1763	1708	1245
(95)	5-HT-1b	Serotonin 1b Rezeptor	GPCR	aminerg		K_i	1024	635	582	360
(96)	5-HT-1d	Serotonin 1d Rezeptor	GPCR	aminerg		K_i	350	256	211	134
(97)	5-HT-2a	Serotonin 2a Rezeptor	GPCR	aminerg		IC_{50}	886	659	395	271
(98)	5-HT-2b	Serotonin 2b Rezeptor	GPCR	aminerg		K_i	853	559	511	336
(99)	5-HT-2c	Serotonin 2c Rezeptor	GPCR	aminerg		K_i	1761	1103	903	602

(100)	5-HT-6	Serotonin 6 Rezeptor	GPCR	aminerg		K_i	2082	1580	1123	877
(101)	M1	Muskarinischer Acetylcholin 1 Rezeptor	GPCR	aminerg		K_i	1591	843	688	465
(102)	M2	Muskarinischer Acetylcholin 2 Rezeptor	GPCR	aminerg		K_i	1793	914	830	601
(103)	M3	Muskarinischer Acetylcholin 3 Rezeptor	GPCR	aminerg		K_i	1116	670	594	418
(104)	M4	Muskarinischer Acetylcholin 4 Rezeptor	GPCR	aminerg		K_i	611	228	223	155
(105)	AT-1	Angiotensin II Typ 1 Rezeptor	GPCR	peptidisch		IC_{50}	1972	1422	1184	892
(106)	AT-2	Angiotensin II Typ 2 Rezeptor	GPCR	peptidisch		IC_{50}	1064	609	487	410
(107)	B2	Bradykinin-Rezeptor 2	GPCR	peptidisch		K_i	589	426	350	232
(108)	CCK-A	Cholecystokinin A Rezeptor (CCK-1)	GPCR	peptidisch		IC_{50}	1274	749	833	563
(109)	CCR2	Chemokin-Rezeptor 1	GPCR	peptidisch		IC_{50}	2059	1107	957	697
(110)	CCR3	Chemokin-Rezeptor 3	GPCR	peptidisch		IC_{50}	953	430	544	438
(111)	CCR5	Chemokin-Rezeptor 5	GPCR	peptidisch		IC_{50}	2260	1500	1448	993
(112)	ET-alpha	Endothelin-Rezeptor alpha	GPCR	peptidisch		IC_{50}	2613	1679	1449	1036
(113)	ET-beta	Endothelin-Rezeptor beta	GPCR	peptidisch		IC_{50}	1815	936	866	608
(114)	Ghrelin-1	Ghrelin-Rezeptor 1 (growth	GPCR	peptidisch		IC_{50}	840	498	453	336

		hormone secretagogue receptor 1)								
(115)	Gonatropin-RHR	Gonadotropin-Releasing Hormon Rezeptor (GNRHR)	GPCR	peptidisch		IC_{50}	2267	527	633	479
(116)	Melanocortin-3	Melanocortin-Rezeptor 3 (MC3R)	GPCR	peptidisch		K_i	602	247	371	262
(117)	NK-1	Neurokinin-Rezeptor 1 (Tachykinin-Rezeptor 1)	GPCR	peptidisch		IC_{50}	2300	1567	1957	1188
(118)	Opioid-delta	Delta opioid receptor	GPCR	peptidisch		IC_{50}	1524	579	999	765
(119)	Opioid-kappa	Kappa opioid receptor	GPCR	peptidisch		K_i	4677	2605	3954	3072
(120)	Opioid-mu	Mu opioid receptor	GPCR	peptidisch		K_i	5565	3267	4923	3611
(121)	Orexin-1	Orexin-Rezeptor 1 (Hypocretin-Rezeptor 1, OX-1)	GPCR	peptidisch		IC_{50}	789	379	569	331
(122)	Orexin-2	Orexin-Rezeptor 2 (Hypocretin-Rezeptor 2, OX-2)	GPCR	peptidisch		IC_{50}	1037	389	571	339
(123)	Somatostatin-5	Somatostatin-Rezeptor 5 (SST-5)	GPCR	peptidisch		K_i	626	145	133	115
(124)	mGlut-1	Metabotroper Glutamat Rezeptor-1	GPCR	Glutamat		IC_{50}	787	483	352	243

(125)	mGlut-5	Metabotroper Glutamat Rezeptor-1	GPCR	Glutamat		IC_{50}	1822	1125	579	422
(126)	Melatonin-1B	Melatonin-Rezeptor 1B (MT-2, MTNR1B)	GPCR	Melatonin		K_i	1366	377	218	171
(127)	P2Y12	P2Y-12-Rezeptor (ADP-Rezeptor)	GPCR	purinerg		K_i	702	558	222	188
(128)	CB-1	Cannabinoid Rezeptor-1	GPCR	lipidisch		K_i	3243	1939	1604	1221
(129)	CB-2	Cannabinoid Rezeptor-2	GPCR	lipidisch		K_i	2325	1597	1579	1180
(130)	PG-D2	Prostaglandin D2 Rezeptor (DP1)	GPCR	lipidisch		IC_{50}	1023	510	451	265
(131)	Glutamat-Kainat	Glutamat-Rezeptor Kainat	Ionen-kanal	Liganden-gesteuert		IC_{50}	549	79	59	54
(132)	Glutamat-NMDA	Glutamat-Rezeptor NMDA	Ionen-kanal	Liganden-gesteuert		IC_{50}	1106	372	285	210
(133)	5-HT-3A	Serotonin-3A-Rezeptor	Ionen-kanal	Liganden-gesteuert		K_i	314	235	167	116
(134)	nAChR	Nikotinischer Acetylcholin-Rezeptor (Neuronal) (Alpha7)	Ionen-kanal	Liganden-gesteuert		K_i	1364	726	274	194
(135)	P2X7	P2X-Purinorezeptor 7	Ionen-kanal	Liganden-gesteuert		IC_{50}	1201	668	598	402
(136)	Sigma-Opioid	Sigma non-opioid intracellular receptor 1	Orphan	(ligand-gesteurtes ER-Chaperon)		K_i	1939	1295	722	573

(137)	DAT	Dopamin Transporter	Trans-porter			K_i	3696	2031	1140	915
(138)	NET	Noradrenalin Transporter	Trans-porter			K_i	2937	1788	913	711
(139)	SERT	Serotonin Transporter	Trans-porter			K_i	5191	3259	1893	1432
(140)	SLGT2	Sodium-dependent glucose cotransporter 2	Trans-porter			IC_{50}	715	561	315	252

III. Ergebnisse und Diskussion

18. Ergebnisse und Diskussion: Netzwerk-Optimierung und Ähnlichkeits-Analyse

Im Folgenden werden die Ergebnisse der Analyse der Variation einiger optionalen Parameter bei der inSARa-Netzwerk-Erstellung gezeigt und ihr Einfluss auf die Komplexität und Spezifität der resultierenden Netzwerke diskutiert.

18.1. Identifizierung unspezifischer RG-MCSs

Zur allgemeinen Analyse der Spezifität von RG-MCSs und zur Analyse der Spezifität der festgelegten Mindest-MCS-Größe von 3 Pseudoatomen wurde der unter 13.1 beschriebene Versuchsaufbau verwendet. Hierbei wurde die Ähnlichkeit von Zufallspaaren aus der ZINC-Datenbank bestimmt. Die Ergebnisse dieser Analyse sind in den Abbildung 18.1, Abbildung 18.2 und Abbildung 18.3 zusammengefasst.

Analyse der RG-MCS-Größe von Zufalls-Molekülpaaren

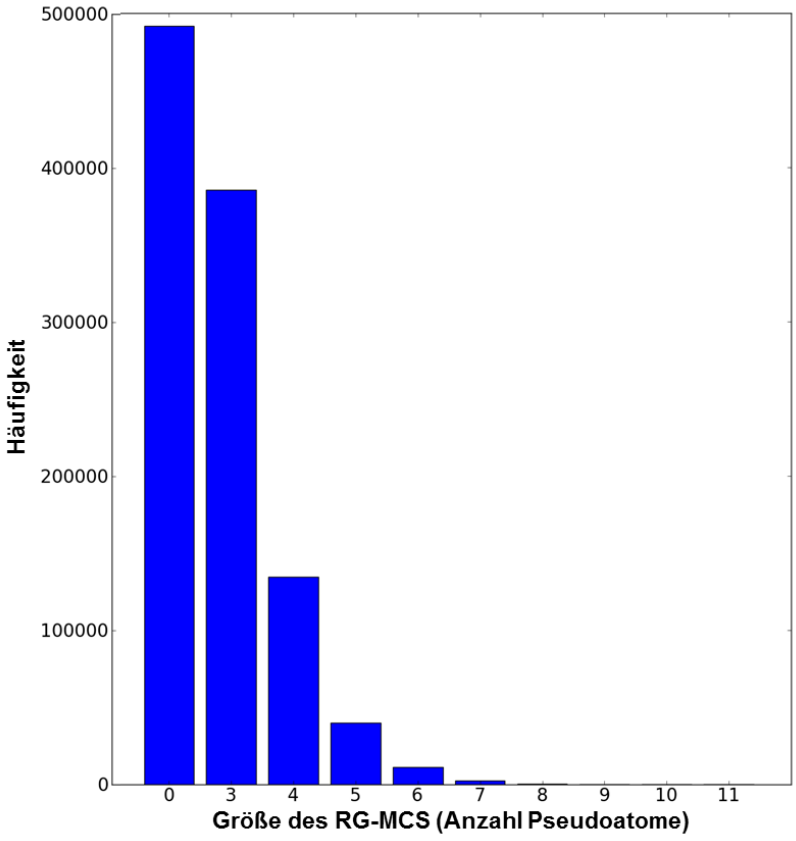


Abbildung 18.1. Häufigkeit des Auftretens der verschiedenen RG-MCS-Größen beim Vergleich von einer Millionen Zufalls-Molekülpaaren aus der ZINC Datenbank.

In Abbildung 18.1 wird die Häufigkeit verschiedener RG-MCS-Größen bei der Analyse von einer Millionen Zufallspaaren gezeigt. Man sieht, dass fast fünfzig Prozent der Zufalls-Paare keinen RG-MCS von mindestens 3 Pseudoatomen aufweisen. Das bedeutet, dass die festgelegte Mindest-MCS-Größe, die bei der Netzwerk-Erzeugung verwendet wird, bereits eine recht hohe Spezifität aufweist. Erwartungsgemäß nimmt die Häufigkeit von Zufalls-RG-MCSs sehr stark mit zunehmender MCS-Größe ab. MCSs, die aus 6 Pseudoatomen bestehen, treten nur noch selten (< 1%) in Zufallspaaren auf. Aus diesen Ergebnissen lässt sich schließen, dass eine Erhöhung der Mindest-MCS-Größe auf 4 oder 5 Pseudoatome dazu beitragen kann, MCSs, die auf Zufalls-Ähnlichkeit beruhen, von inSARa-Netzwerken auszuschließen. Dadurch ist es wahrscheinlicher, dass die MCSs, die im Netzwerk gezeigt werden, Zielstruktur-spezifische gemeinsame Merkmale repräsentieren.

Identifizierung einzelner unspezifischer RG-MCSs zur Erstellung einer Ausschlussliste

Die Häufigkeit einzelner RG-MCSs wurde ebenfalls in den gezogenen Zufalls-Molekülpaaren analysiert. Abbildung 18.2 zeigt die Häufigkeit der 20 am häufigsten auftretenden RG-MCSs. Die RG-MCSs sind jeweils als SMILES-Strings dargestellt.

Die Ergebnisse dieser Analyse können ebenfalls dazu verwendet werden, die Zielstruktur-Spezifität der inSARa-Netzwerke zu erhöhen. Zu diesem Zweck wird eine Zufallswahrscheinlichkeit von 0,1% als Grenze für mangelnde Spezifität festgelegt. Die Wahrscheinlichkeit berechnet sich aus dem Quotienten der jeweiligen Häufigkeit und der Gesamtzahl an Zufallspaaren (= 10^6). Eine komplette Liste aller RG-MCSs mit einer Auftretens-Wahrscheinlichkeit größer oder gleich 0,1% findet sich in Tabelle 26.3 im Anhang (Abschnitt 26.3). Diese Liste wird im Folgenden als „Ausschlussliste“ bezeichnet. Die Grenze von 0,1% wurde empirisch nach visueller Inspektion dieser MCSs festgelegt unter Sicherstellung, dass keine spezifische Information codiert ist.

Wenn die Ausschlussliste zum Ausschluss von MCSs aus inSARa-Netzwerken verwendet wird, wird dies im Folgenden kurz mit „Ausschlussliste = aktiv“ bezeichnet. Standardmäßig wird diese Liste aktiv gesetzt, optional kann sie jedoch vom Nutzer deaktiviert werden. Alle RG-MCSs mit einer Wahrscheinlichkeit größer oder gleich 0,1% werden somit standardmäßig als unspezifisch betrachtet und von der Gesamtmenge an einzigartigen MCSs verworfen (vgl. Abschnitt 10.3). Dadurch werden sie nicht weiter bei der Netzwerk-Erzeugung berücksichtigt.

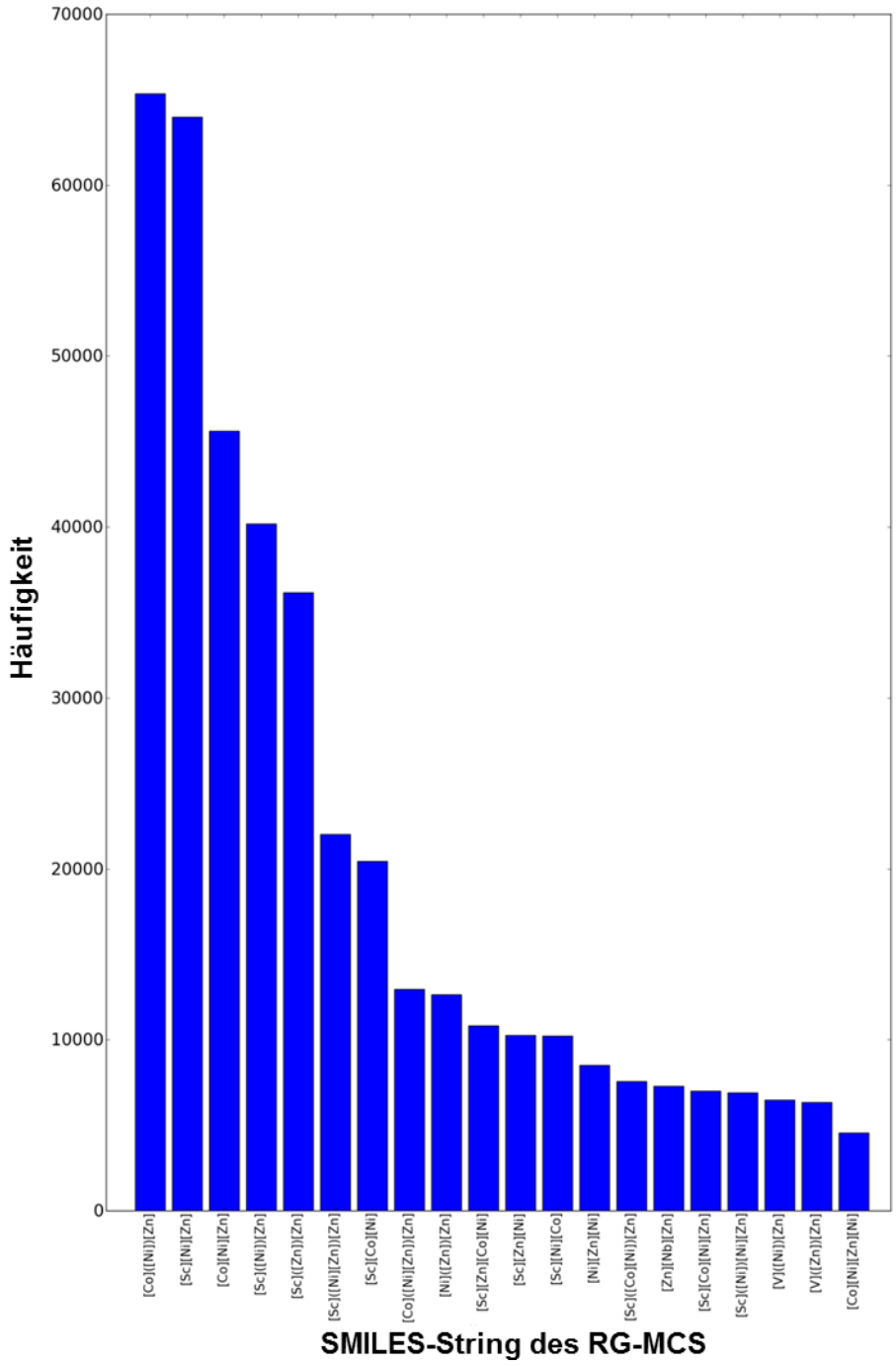


Abbildung 18.2. Häufigkeit der 20 häufigsten RG-MCSs (dargestellt als SMILES-String) beim Vergleich von einer Millionen Zufalls-Molekülpaaren aus der ZINC Datenbank. Falls mehrere gleich große RG-MCSs für ein Molekülpaar bestimmt werden, werden alle MCSs in dieser Statistik berücksichtigt.

Bei Verwendung des Schwellenwertes von 0,1% umfasst die Ausschlussliste 60 MCSs der Größe von 3 oder 4 Pseudoatomen. Die häufigsten Pseudoatome sind Sc (= aromatischer Ring ohne weitere Eigenschaften), Zn (= Linker oder terminale Gruppe), Ni (= nicht in ein Ringsystem eingebundener HBA) und Co (= azyklischer HBD). Der häufigste SMILES [Co]([Ni])[Zn] repräsentiert einen Linker, der mit einem HBD verknüpft ist, der wiederum zu einem HBA benachbart ist. Die Kombination aus Co und Ni repräsentiert zum Beispiel Sulfonamide oder Carbonsäureamide, die typischerweise Bestandteil vieler peptidartiger

Moleküle sind. Diese Kombination ist somit wenig spezifisch. Sc repräsentiert beispielsweise Phenylringe. Diese stellen eine der häufigsten sterischen Komponenten in Arzneistoff-ähnlichen Molekülen dar. Sc ist somit ebenso unspezifisch wie Zn, das jeglich Linkeratome oder terminale Gruppen codiert.

Zusammenfassend lässt sich somit feststellen, dass bei Verwendung dieser Ausschlussliste damit zu rechnen ist, dass wenig SAR-Information verloren geht und die Spezifität der Netzwerke gesteigert werden kann. Da nur kleine MCSs Bestandteil dieser Liste sind, spielt diese Liste ab einer Mindest-MCS-Größe von 5 RG-Atomen jedoch keine Rolle mehr. Bei einer Mindest-MCS-Größe von 3 oder 4 Pseudoatomen, ist bei der Analyse von großen Datensätzen aufgrund der geringen Zahl an unspezifischen MCSs im Vergleich zur Gesamtzahl an MCSs nur eine geringe Abnahme der Netzwerk-Komplexität zu beobachten.

18.2. Vergleich von Fingerprint- und MCS-basierter Ähnlichkeit

Vergleich von Fingerprint-Ähnlichkeit und MCS-Größe in Zufallspaaren

In Abbildung 18.3 ist die Fingerprint-Ähnlichkeit (ECFP4 und MACCS Keys) der Zufalls-Molekülpaare in Abhängigkeit von der RG-MCS-Größe in Form von Boxplots dargestellt. Es ist hierbei, wie zu erwarten, eine leichte Zunahme der FP-Ähnlichkeit mit zunehmender MCS-Größe zu erkennen. Es ist jedoch zu beachten, dass die Anzahl an Molekülpaaren für alle MCS-Größen (wie in Abbildung 18.1 dargestellt) nicht gleich ist, sondern für größere MCSs die Anzahl sehr gering ist, sodass der dargestellte Trend ggf. durch die unterschiedlichen Fallzahlen verzerrt sein kann und ein zu beobachtender Unterschied nicht unbedingt statistisch signifikant ist.

In Abbildung 18.3 ist ebenfalls zu erkennen, dass die Tc-Ähnlichkeit bei Verwendung von MACCS Keys deutlich größer ist als beim ECFP4-Fingerprint. Während man bei MACCS Keys als Bereich für erkennbare Ähnlichkeit für gewöhnlich einen Tc von etwa 0.65 bis 0.8 annimmt, liegt dieser Bereich bei dem ECFP4-FP bei etwa 0.4 bis 0.55. Auch wenn der Median bzw. das obere und untere Quartil, v.a. für kleine MCSs, deutlich von diesen Bereichen entfernt sind, ist jedoch bei den MACCS Keys zu beobachten, dass neben einzelnen Ausreißern, die in geringerer Zahl auch bei dem ECFP4-FP zu beobachten sind, der obere Whisker auch für sehr kleine RG-MCSs in diesen Bereich liegt. Eine Begründung dafür, dass mit MACCS Keys trotz kleiner gemeinsamer Substruktur eine relativ große FP-Ähnlichkeit gefunden wird, ist, dass es sich hierbei um einen Wörterbuch-basierten Substruktur-Fingerprint handelt. Hierbei wird nur geprüft, ob bestimmte, vordefinierte Substrukturen in den zu vergleichenden Zufallsmolekülen vorkommen. Im Gegensatz zum ECFP, wo die gesamte Substruktur-Umgebung in einem bestimmten Radium um ein Atom codiert wird und somit auch die Konnektivität codiert bzw. bei der Bestimmung der Ähnlichkeit berücksichtigt wird, wird bei den MACCS Keys die Konnektivität nicht berücksichtigt. Die Wahrscheinlichkeit ist somit relativ groß, dass in Zufalls-Moleküle gleiche Substrukturen gefunden werden, sodass trotz geringer lokaler Ähnlichkeit (kleiner MCS) hohe Tc-Werte resultieren. Ein Chemiker würde diese Moleküle aufgrund der andersartigen Verknüpfung der Substrukturen zumeist nicht als ähnlich einstufen. Beim ECFP-FP ist die Wahrscheinlichkeit für lokale Ähnlichkeit bei hohem Tc deutlich höher.

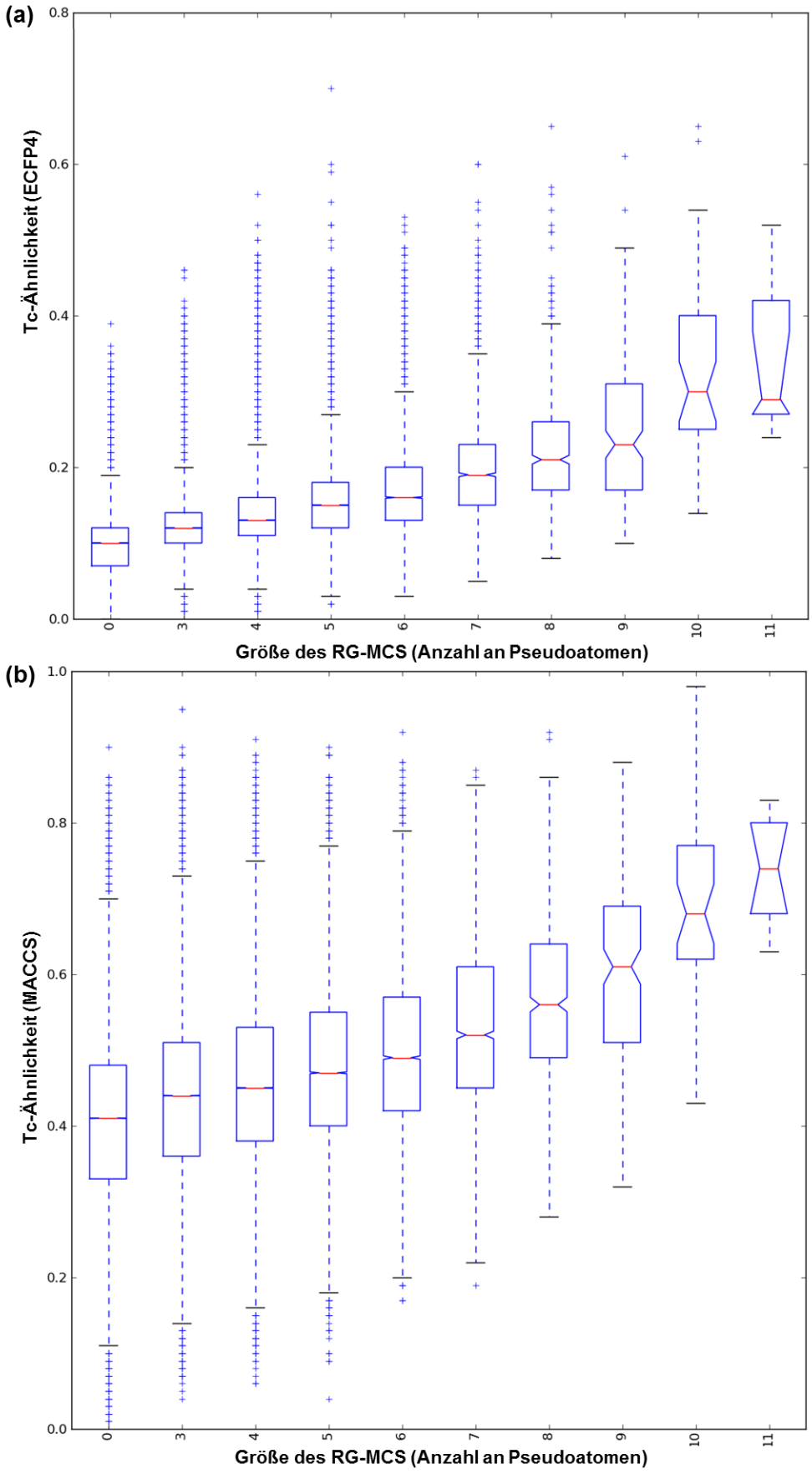


Abbildung 18.3. Zusammenhang zwischen Fingerprint-Ähnlichkeit (Tc) und der Größe der RG-MCSs bei der Analyse von einer Millionen Zufalls-Molekülpaaren aus der ZINC Datenbank. (a) ECFP4-Fingerprint, (b) MACCS Keys.

Vergleich von Fingerprint-Ähnlichkeit und MCS-Größe in Datensatz-Molekülpaaren

In Abbildung 18.4 ist in Analogie zu Abbildung 18.1 die Häufigkeit verschiedener RG-MCS-Größen beim paarweisen Vergleich aller Moleküle verschiedener Datensätze dargestellt. In Abbildung 18.5 ist in Analogie zu Abbildung 18.3 die Tc-basierte Fingerprint-Ähnlichkeit (ECFP4 und MACCS Keys) in Abhängigkeit von den RG-MCS-Größen dargestellt.

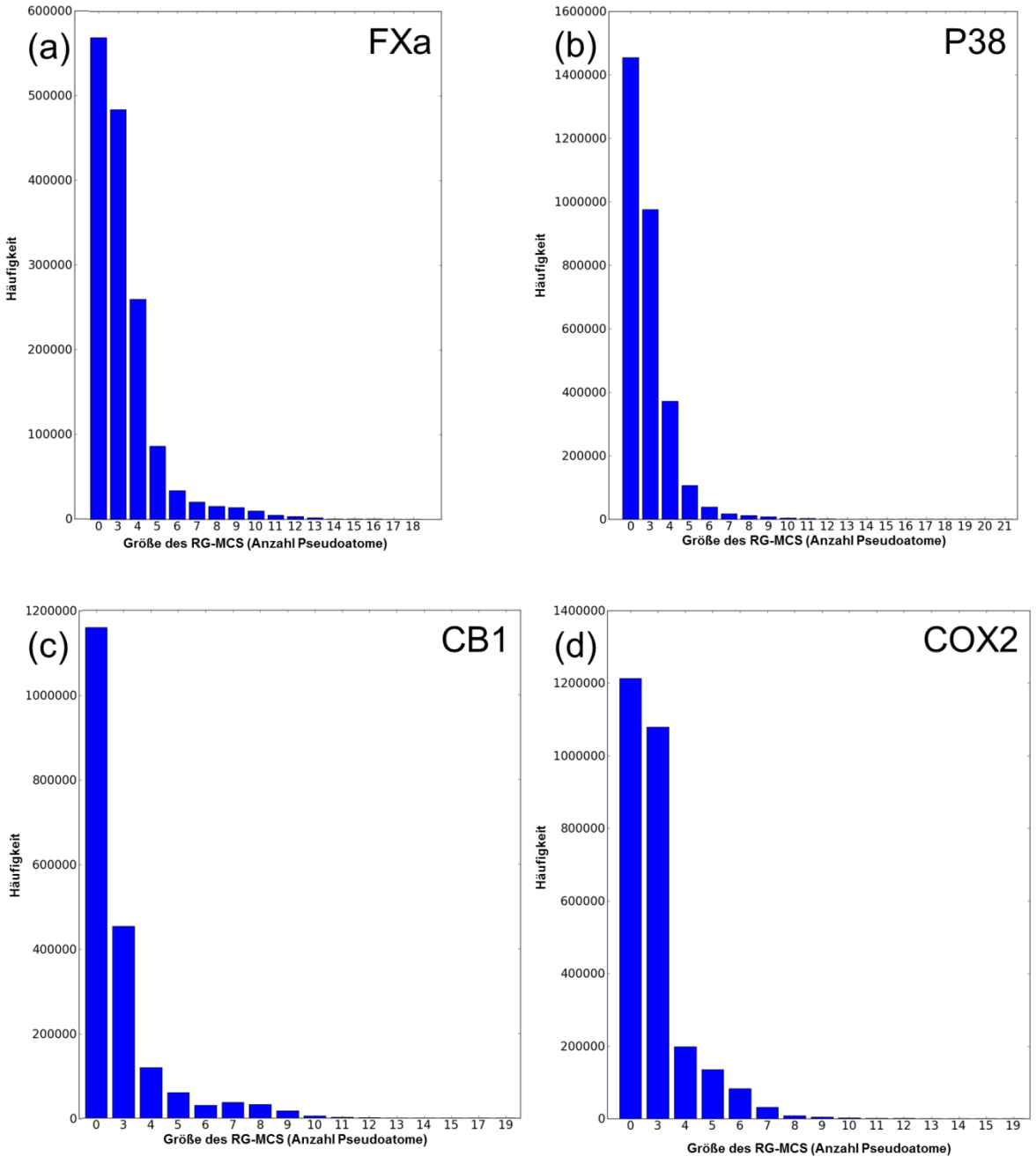


Abbildung 18.4. Häufigkeit des Auftretens der verschiedenen RG-MCS-Größen in den Molekülpaaren verschiedener Datensätze: (a) FXa, (b) P38, (c) CB1, (d) COX2.

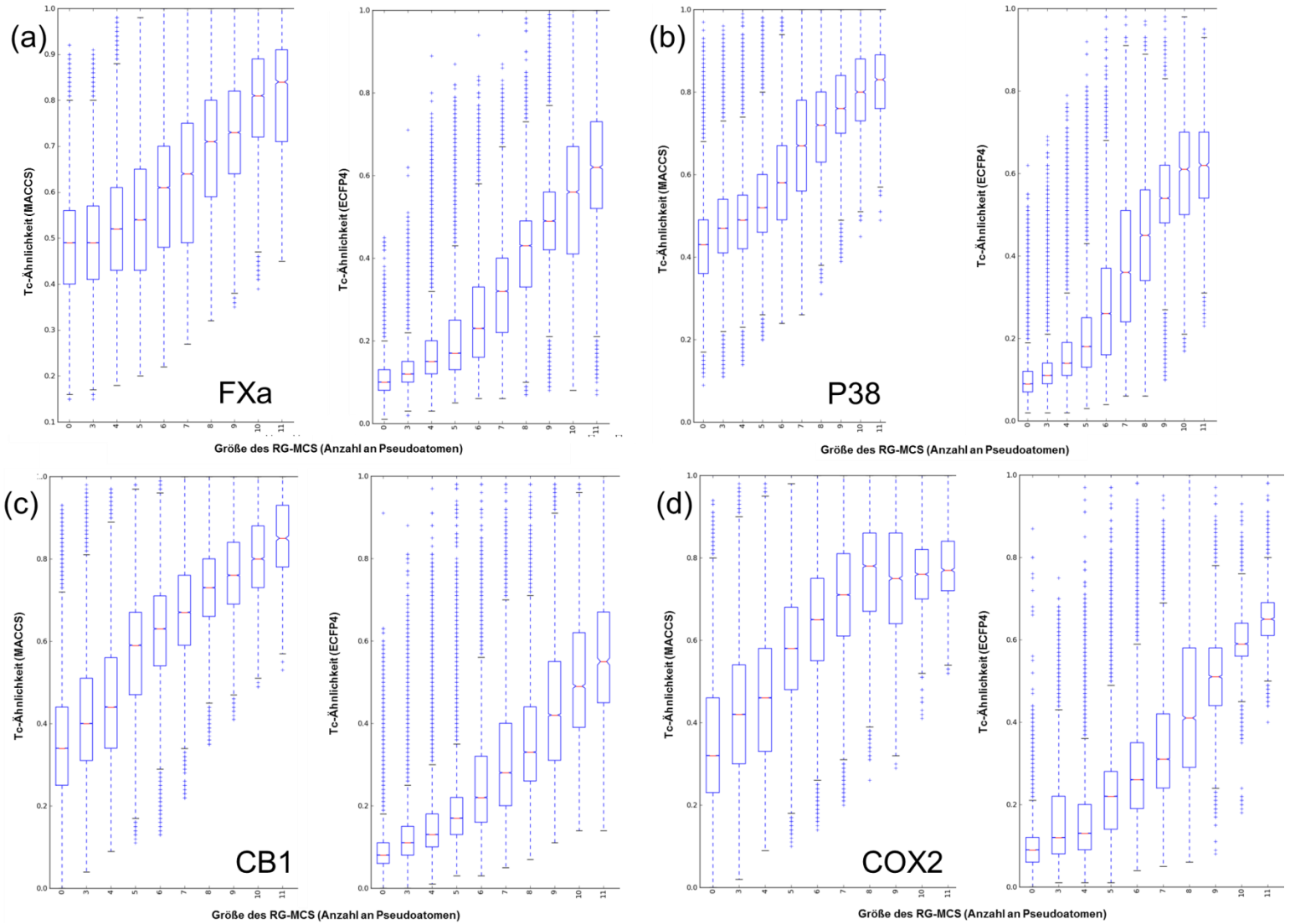


Abbildung 18.5. Zusammenhang zwischen Fingerprint-Ähnlichkeit (Tc) für MACCS Keys und den ECFP4-Fingerprint und der Größe der RG-MCSs in den Molekülpaaren verschiedener Datensätze: (a) FXa, (b) P38, (c) CB1, (d) COX2. MCSs > 11 RG-Atome sind aufgrund geringer Häufigkeit nicht dargestellt.

In Abbildung 18.4 ist zu sehen, dass auch in Datensätzen, wo eine Beziehung zwischen den Molekülen bekannt ist, ein sehr großer Anteil der Molekülpaare keinen RG-MCS der Mindest-Größe von 3 Pseudoatomen aufweist. MCS-Größen ab 5 oder 6 RG-Atomen sind auch hier im Vergleich zur Gesamtzahl an Molekülpaaren relativ selten.

In Abbildung 18.5 zeigt sich ebenfalls der in Abbildung 18.3 beobachtete Trend zwischen Fingerprint-Ähnlichkeit und MCS-Größe. Auch hier ist wieder zu beachten, dass für größere MCSs die Zahl an verfügbaren Ähnlichkeitswerten immer kleiner wird und so der Vergleich erschwert wird. Auffällig sind die Ausreißer, die einen hohen ECFP4-Tc-Ähnlichkeitswert bei kleinem RG-MCS aufweisen. Dies könnte beispielsweise durch suboptimale RG-Codierung begründet sein. Es könnte aber auch ein Hinweis auf das Vorhandensein einer größeren nicht-zusammenhängenden gemeinsamen Substruktur sein, die hierbei jedoch nicht berücksichtigt wird.

Korrelation zwischen Fingerprint-Ähnlichkeit und RG-MCS-Ähnlichkeit

In Abbildung 18.6 ist die Korrelation (Verwendung des Spearman-Rang-Korrelationskoeffizienten) der Fingerprint-Ähnlichkeit und MCS-Ähnlichkeit beispielhaft anhand des FXa-Datensatzes dargestellt. Für die anderen Datensätze sind die Korrelations-Daten jedoch vergleichbar. Zwischen RASCAL-Ähnlichkeit, die die MCS-Größe ins Verhältnis zur Größe der RGs der Moleküle setzt, und der MCS-Größe zeigt sich eine fast perfekte positive Korrelation. Wie schon zuvor gesehen, besteht eine stärkere positive Korrelation zwischen dem ECFP4-Fingerprint und der RG-MCS-Ähnlichkeit als zwischen den MACCS Keys und allen anderen Fingerprints. Der Pfad-basierte FP2-Fingerprint weist ebenfalls eine höhere Korrelation als die MACCS Keys zu RG-MCS-basierter Ähnlichkeit auf. Dies könnte ebenfalls auf die Berücksichtigung von Konnektivität durch die Kodierung von Molekülpfaden zurückzuführen sein. Die Codierung der Häufigkeit des Vorkommens von Substrukturen (Vergleich MACCSF und MACCS) führt ebenfalls zu einer erhöhten Korrelation. Zusammenfassend lässt sich feststellen, dass zwar eine gewisse Korrelation zwischen Fingerprint-Ähnlichkeit (v.a. ECFP4-basiert) und RG-MCS-Ähnlichkeit vorhanden ist. Jedoch kann man auch sehen, dass deutliche Unterschiede durch die verschiedenen Arten der Erfassung von molekularer Ähnlichkeit (verschiedene Formen der molekularen Repräsentationen und Ähnlichkeitsmetriken) zu beobachten sind, sodass auch für die SAR-Analyse zu erwarten ist, dass molekulare Beziehungen anders erfasst werden und andere SAR-Informationen gewonnen werden können.

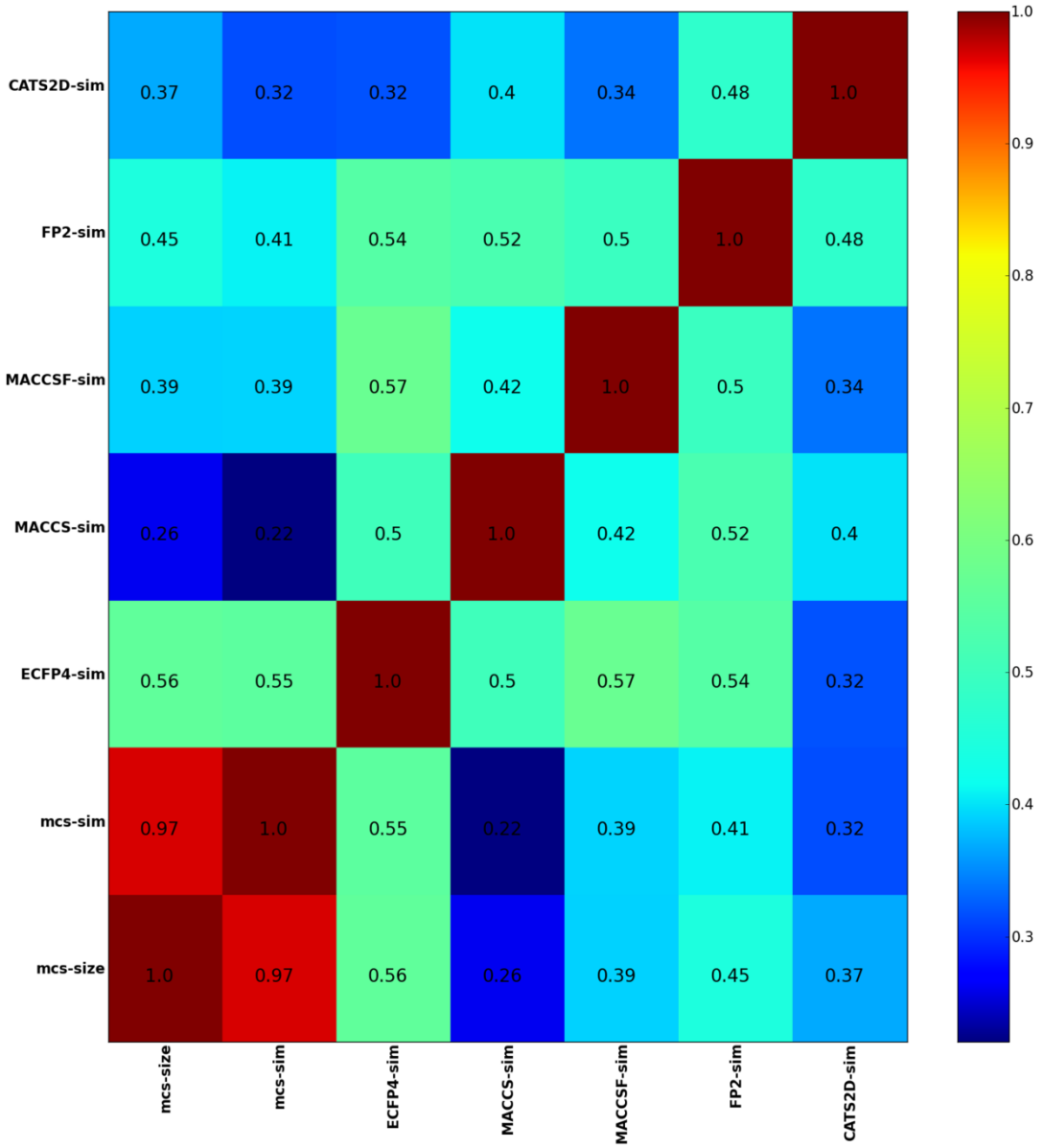


Abbildung 18.6. Vergleich der Korrelation von Fingerprint- und RG-MCS-Ähnlichkeit anhand des Spearman-Rang-Korrelationskoeffizienten für Molekülpaare im FXa-Datensatz. MCS size = RG-MCS-Größe, MCS-sim = RG-MCS-basierter RASCAL-Score, Rest = Tc-Ähnlichkeit verschiedener Fingerprints.

18.3. Weitere optionale Optimierungs-Parameter

Zusätzlich zur (In-)Aktivierung der Ausschlussliste, stellen die Mindest-MCS-Größe und das Abbruchkriterium für die Wurzel-Knoten-Auswahl zwei weitere zusätzliche Parameter zum Fine-Tuning der Netzwerke dar, die von dem Benutzer variiert werden können. Die Ergebnisse der unter 13.3 beschriebenen Analyse sind in der nachfolgenden Tabelle 18.1 und Tabelle 18.3 zusammengefasst und werden in im Folgenden diskutiert.

18.3.1. Variation der Mindest-MCS-Größe

Bei der Netzwerk-Erzeugung ist die Mindest-MCS-Größe auf 3 RG-Atome festgelegt. RG-MCSs, die aus 3 RG-Pseudoatomen bestehen, sind jedoch in vielen Fällen wenig spezifisch. Dies belegen auch die oben gezeigten Ergebnisse der Analyse der ZINC-Datenbank (siehe Abschnitt 18.1). Hier wird gezeigt, dass RG-MCSs der Größe von 3 RG-Atomen mit einer Häufigkeit von etwa 40% (vgl. Abbildung 18.1) in Molekülen vorkommen, die in keiner bekannten Beziehung zueinander stehen. Durch Erhöhung des Wertes für die Mindest-MCS-Größe wird die Anzahl an MCSs und nachfolgend auch die Komplexität des resultierenden Netzwerkes reduziert (vgl. Tabelle 18.1). In den meisten Fällen führt ein Ausschluss kleinerer RG-MCSs nur zu einem marginalen Informationsverlust. Denn diese MCSs repräsentieren oftmals weitverbreitete Merkmale mit wenig Informationsgehalt für die SAR-Analyse. Sie haben häufig nur die Funktion, die größeren MCSs in dem Netzwerk zu verbinden. Daher führt die Erhöhung der Mindest-MCS-Größe auch zu einer höheren Anzahl an nicht-zusammenhängenden Komponenten in den inSARa-Netzwerken (vgl. Tabelle 18.1). Nichtdestotrotz bleibt die wichtige Information (für gewöhnlich wird diese an MCS-Knoten der Größe 8 und größer gefunden) erhalten. Ein Haupt-Vorteil ist, dass diese Netzwerke weniger komplex sind durch das vereinfachte Layout und auch schneller zu berechnen sind. Erwartungsgemäß steigt die Anzahl an Wurzel-Knoten mit größerer Mindest-MCS-Größe an (vgl. Tabelle 18.1). Dies kann dadurch erklärt werden, dass diese größeren Wurzel-Knoten spezifischer sind und somit weniger Moleküle repräsentieren. Ein Problem, dass bei der Erhöhung der Mindest-MCS-Größe auftreten kann, ist, dass ein höherer Anteil an Molekülen nicht repräsentiert werden kann und die Wurzel-Knoten-Auswahl aufhört, bevor das festgelegte Abbruchkriterium erfüllt ist (z.B. COX2 in Tabelle 18.1). Es wird demzufolge ein Kompromiss benötigt zwischen der Spezifität der Netzwerke und dem Grad der Abdeckung des chemischen Raumes durch die Netzwerke. Denn eine Zunahme der Mindest-MCS-Größe führt zu spezifischeren, aber auch restriktiveren Netzwerken, weil die gleichen Moleküle nicht mehr mit größerer Mindest-MCS-Größe repräsentiert werden können. Abgesehen von der Spezifität und dem Anteil an chemischen Raum, der durch das Netzwerk repräsentiert werden soll, ist auch die Gesamt-Diversität des Datensatzes ein wichtiges Kriterium für die Wahl der idealen Mindest-MCS-Größe. Denn diverse Datensätze sind meist durch kleine, unspezifische MCSs charakterisiert. Ein dritter Faktor für eine gute Mindest-MCS-Größe ist die durchschnittliche RG-Größe der Datensatz-Moleküle (vgl. Tabelle 18.2).

Wie in Tabelle 18.1 zu sehen, führt eine Erhöhung der Mindest-MCS-Größe in Datensätzen mit kleinerer durchschnittlicher RG-Größe (z.B. COX2 und CB1) zu einem höheren Anteil an Molekülen, die nicht durch MCS-Knoten des resultierenden Netzwerkes repräsentiert werden können. Zusammenfassend lässt sich feststellen, dass eine Mindest-MCS-Größe von 5 RG-Atomen in vielen Fällen einen guten Kompromiss aus potentiellem Verlust an SAR-Information und Netzwerk-Komplexität darstellt und somit als Standard-Wert empfehlenswert ist.

Tabelle 18.1. Einfluss der Mindest-MCS-Größe auf die Netzwerk-Komplexität und -Topologie (Abbruch-Kriterium = 2% nicht-repräsentierte Moleküle, Ausschlussliste = aktiv)

Target	Mindest-MCS-Größe	3	4	5	6	7	8
FXa	Einzigartige MCSs in MCS-Matrix	1774	1680	1470	1227	996	804
	Wurzel-Knoten im Netzwerk	12	19	36	57	81	88
	MCS-Knoten im Netzwerk	1041	800	660	541	454	366
	Nicht-repräsentierte Moleküle	33	33	33	34	74	156
	Repräsentierte Moleküle	1703	1703	1703	1702	1662	1580
	Nicht-repräsentierter Anteil (%)	1.9	1.9	1.9	2.0	4.3	9.0
	Anzahl an Komponenten	1	4	16	40	66	74
CDK2	Einzigartige MCSs in MCS-Matrix	1489	1340	1074	795	581	393
	Wurzel-Knoten im Netzwerk	26	53	105	129	150	153
	MCS-Knoten im Netzwerk	986	813	676	488	341	205
	Nicht-repräsentierte Moleküle	31	31	31	127	305	527
	Repräsentierte Moleküle	1544	1544	1544	1448	1270	1048
	Nicht-repräsentierter Anteil (%)	2.0	2.0	2.0	8.1	19.4	33.7
	Anzahl an Komponenten	3	30	76	101	126	146
COX2	Einzigartige MCSs in MCS-Matrix	1750	1615	1322	974	639	366
	Wurzel-Knoten im Netzwerk	31	61	134	190	207	174
	MCS-Knoten im Netzwerk	1336	1117	830	552	315	143
	Nicht-repräsentierte Moleküle	44	45	76	208	639	1261
	Repräsentierte Moleküle	2305	2304	2273	2141	1710	1088
	Nicht-repräsentierter Anteil (%)	1.9	1.9	3.2	8.9	27.2	53.7
	Anzahl an Komponenten	10	29	80	144	171	164
CB1	Einzigartige MCSs in MCS-Matrix	1623	1506	1254	938	669	452
	Wurzel-Knoten im Netzwerk	49	74	96	129	155	158
	MCS-Knoten im Netzwerk	1190	968	761	584	387	233
	Nicht-repräsentierte Moleküle	46	51	88	197	404	643
	Repräsentierte Moleküle	1911	1906	1869	1760	1553	1314
	Nicht-repräsentierter Anteil (%)	2.4	2.6	4.5	10.1	20.6	32.9
	Anzahl an Komponenten	18	41	67	99	127	142

Target	Mindest-MCS-Größe	3	4	5	6	7	8
P38	Einzigartige MCSs in MCS-Matrix	2521	2354	1980	1542	1164	847
	Wurzel-Knoten im Netzwerk	25	43	119	170	219	232
	MCS-Knoten im Netzwerk	1731	1540	1225	992	735	506
	Nicht-repräsentierte Moleküle	48	48	48	141	314	608
	Repräsentierte Moleküle	2398	2398	2398	2305	2132	1838
	Nicht-repräsentierter Anteil (%)	2.0	2.0	2.0	5.8	12.8	24.9
	Anzahl an Komponenten	1	10	67	102	164	183
THR	Einzigartige MCSs in MCS-Matrix	3963	3821	3484	2960	2441	2000
	Wurzel-Knoten im Netzwerk	24	45	105	160	200	206
	MCS-Knoten im Netzwerk	2999	2736	2484	2087	1689	1329
	Nicht-repräsentierte Moleküle	45	56	57	102	217	374
	Repräsentierte Moleküle	2807	2796	2795	2750	2635	2478
	Nicht-repräsentierter Anteil (%)	1.6	2.0	2.0	3.6	7.6	13.1
	Anzahl an Komponenten	1	8	44	102	140	170

Tabelle 18.2. Verteilung der RG-Größe in verschiedenen Datensätzen. Histogramme werden in Abbildung 26.1 im Anhang gezeigt.

Target	Minimale RG-Größe	Maximale RG-Größe	Median der RG-Größe	Mean der RG-Größe
FXa	5	19.0	12.0	11.9
THR	1	24	12.0	11.5
CB1	2	19	9.0	9.0
COX2	2	20	8.2	8.0
CDK2	3	19	9.0	9.2
P38	3	21	10.0	9.8

18.3.2. Variation des Abbruch-Kriteriums für die Wurzel-Knoten-Auswahl

Durch die Variation des Abbruch-Kriteriums für die Wurzel-MCS-Auswahl wird die Netzwerk-Komplexität und -Topologie ebenfalls beeinflusst. Ähnlich wie bei der Mindest-MCS-Größe hängt dieser Parameter desgleichen stark von dem analysierten Datensatz ab. Daher besteht zum Fine-Tuning des Netzwerkes die Möglichkeit der optionalen Modifikation. In Tabelle 18.3 ist zu sehen, dass eine Erhöhung des Abbruch-Kriteriums zu einem Ausschluss von Wurzel-Knoten führt, die nur eine kleine Zahl an Molekülen repräsentieren. Gleichsam zur Anzahl an Wurzel-Knoten nimmt auch die Anzahl an MCS-Knoten und die Anzahl an Komponenten ab. Somit können Netzwerke durch das Beenden der Wurzel-Knoten-Auswahl zu früheren Zeitpunkten ebenfalls vereinfacht werden.

Tabelle 18.3. Einfluss des Abbruch-Kriteriums (bei der Wurzel-Knoten-Auswahl) auf die Netzwerk-Komplexität und -Topologie (Ausschlussliste = aktiv)

Target (Mindest-MCS-Größe)	**Abbruch-Kriterium (Anteil nicht-repräsentierter Moleküle in %)**	**1**	**2**	**5**	**7**	**10**	**15**	**20**	**25**
FXa (5)	Wurzel-Knoten im Netzwerk	43	36	28	25	21	17	13	11
	MCS-Knoten im Netzwerk	666	660	640	629	610	575	532	514
	Nicht-repräsentierte Moleküle	17	33	81	111	168	244	339	404
	Repräsentierte Moleküle	1719	1703	1655	1625	1568	1492	1397	1332
	Nicht-repräsentierter Anteil (%)	1.0	1.9	4.7	6.4	9.7	14.1	19.5	23.3
	Anzahl an Komponenten	21	16	13	13	13	11	8	7
CDK2 (5)	Wurzel-Knoten im Netzwerk	105	105	70	59	49	38	31	27
	MCS-Knoten im Netzwerk	676	676	660	646	618	570	531	495
	Nicht-repräsentierte Moleküle	30	31	77	110	155	231	314	384
	Repräsentierte Moleküle	1545	1544	1498	1465	1420	1344	1261	1191
	Nicht-repräsentierter Anteil (%)	1.9	2.0	4.9	7.0	9.8	14.7	20.0	24.4
	Anzahl an Komponenten	74	76	55	47	39	32	26	22
COX2 (3)	Wurzel-Knoten im Netzwerk	44	31	19	16	14	11	9	7
	MCS-Knoten im Netzwerk	1366	1336	1261	1153	1103	1034	985	898
	Nicht-repräsentierte Moleküle	23	44	110	162	215	316	421	586
	Repräsentierte Moleküle	2326	2305	2239	2187	2134	2033	1928	1763
	Nicht-repräsentierter Anteil (%)	1.0	1.9	4.7	6.9	9.2	13.5	17.9	24.9
	Anzahl an Komponenten	17	10	5	4	4	3	2	1
CB1 (3)	Wurzel-Knoten im Netzwerk	49	49	28	22	18	14	11	9
	MCS-Knoten im Netzwerk	1190	1190	1087	1048	1026	969	855	770
	Nicht-repräsentierte Moleküle	46	46	94	133	192	274	376	468
	Repräsentierte Moleküle	1911	1911	1863	1824	1765	1683	1581	1489
	Nicht-repräsentierter Anteil (%)	2.4	2.4	4.8	6.8	9.8	14.0	19.2	23.9
	Anzahl an Komponenten	18	18	10	7	5	3	4	4

Target (Mindest-MCS-Größe)	**Abbruch-Kriterium (Anteil nicht-repräsentierter Moleküle in %)**	**1**	**2**	**5**	**7**	**10**	**15**	**20**	**25**
P38 (5)	Wurzel-Knoten im Netzwerk	126	119	70	60	49	37	29	24
	MCS-Knoten im Netzwerk	1253	1225	1172	1139	1090	115	961	892
	Nicht-repräsentierte Moleküle	40	48	122	167	244	366	485	608
	Repräsentierte Moleküle	2406	2398	2324	2279	2202	2080	1961	1838
	Nicht-repräsentierter Anteil (%)	1.6	2.0	5.0	6.8	10.0	15.0	19.8	24.9
	Anzahl an Komponenten	67	67	32	29	23	16	12	9
THR (6)	Wurzel-Knoten im Netzwerk	160	160	122	96	73	31	37	29
	MCS-Knoten im Netzwerk	2087	2087	2058	2019	1964	1849	1804	1696
	Nicht-repräsentierte Moleküle	102	102	142	198	282	423	557	700
	Repräsentierte Moleküle	2750	2750	2710	2654	2570	2429	2295	2152
	Nicht-repräsentierter Anteil (%)	3.6	3.6	5.0	7.0	9.9	14.8	19.5	24.5
	Anzahl an Komponenten	102	102	81	63	46	33	21	16

18.3.3. Weitere Möglichkeiten der Netzwerk-Modifikation

Eine weitere Möglichkeit des Ausschlusses von Wurzel-Knoten, die nur eine kleine Anzahl an Molekülen repräsentieren, wäre die Festlegung einer Mindest-Zahl an Molekülen, die durch den Wurzel-MCS repräsentiert werden müssen. Wird diese Zahl unterschritten, würde der Wurzel-MCS bei der Wurzel-Knoten-Auswahl nicht dem Netzwerk hinzugefügt werden. Falls keine weiteren Wurzel-Knoten, die dieses Kriterium erfüllen, gefunden werden können, könnte die Wurzel-Knoten-Auswahl gestoppt werden.

Die Netzwerk-Komplexität lässt sich ebenfalls nachträglich durch Zurückschneiden des Netzwerkes reduzieren. Durch die Definition einer Mindest-Anzahl an Molekülen, die durch terminale Knoten repräsentiert werden sollen, können beispielsweise terminale Knoten, die nur wenige Moleküle repräsentieren vom Netzwerk ausgeschlossen werden und die Komplexität folglich reduziert werden. Dies ist insbesondere dann sinnvoll, wenn der zurückgeschnittene Ast ohnehin „einfarbig“ (d.h. einheitliche Bioaktivität) ist. Außerdem ist zu beachten, dass die Anzahl an MCS-Knoten in Tabelle 18.1 und Tabelle 18.3 alle MCS-Knoten des Netzwerkes berücksichtigt, d.h. auch Knoten, an denen keine Moleküle gezeigt werden. Diese MCSs können ebenfalls aus dem Netzwerk entfernt werden, sofern sie keine Verbindungs-Funktion haben. Viele weitere Kriterien zur Vereinfachung des Netzwerkes können definiert werden. Auch wäre es denkbar nur MCS-Knoten, an denen die MCS-Ähnlichkeit (bezogen auf die durchschnittliche MCS-Größe, vgl. Abschnitt 16.2) einen bestimmten Schwellenwert übersteigt im Netzwerk zu behalten, da hier die Wahrscheinlichkeit für Fehlgruppierungen deutlich reduziert ist und für gut-interpretierbare Beziehungen mit hohem Nutzen für die SAR-Interpretation groß ist.

Zusammenfassend ist festzustellen, dass in Abhängigkeit von den Erfordernissen der jeweiligen Analyse verschiedene Möglichkeiten des Netzwerk-Fine-Tunings und der Netzwerk-Vereinfachung möglich sind.

19. Ergebnisse und Diskussion: Vergleich nächster Nachbarn

19.1. Ergebnisse

Die Ergebnisse der in Kapitel 15 beschriebenen *k*NN-Regressions-basierten Bioaktivitäts-Vorhersage sind in Abbildung 19.1 bis Abbildung 19.4 in Form von Boxplots dargestellt.

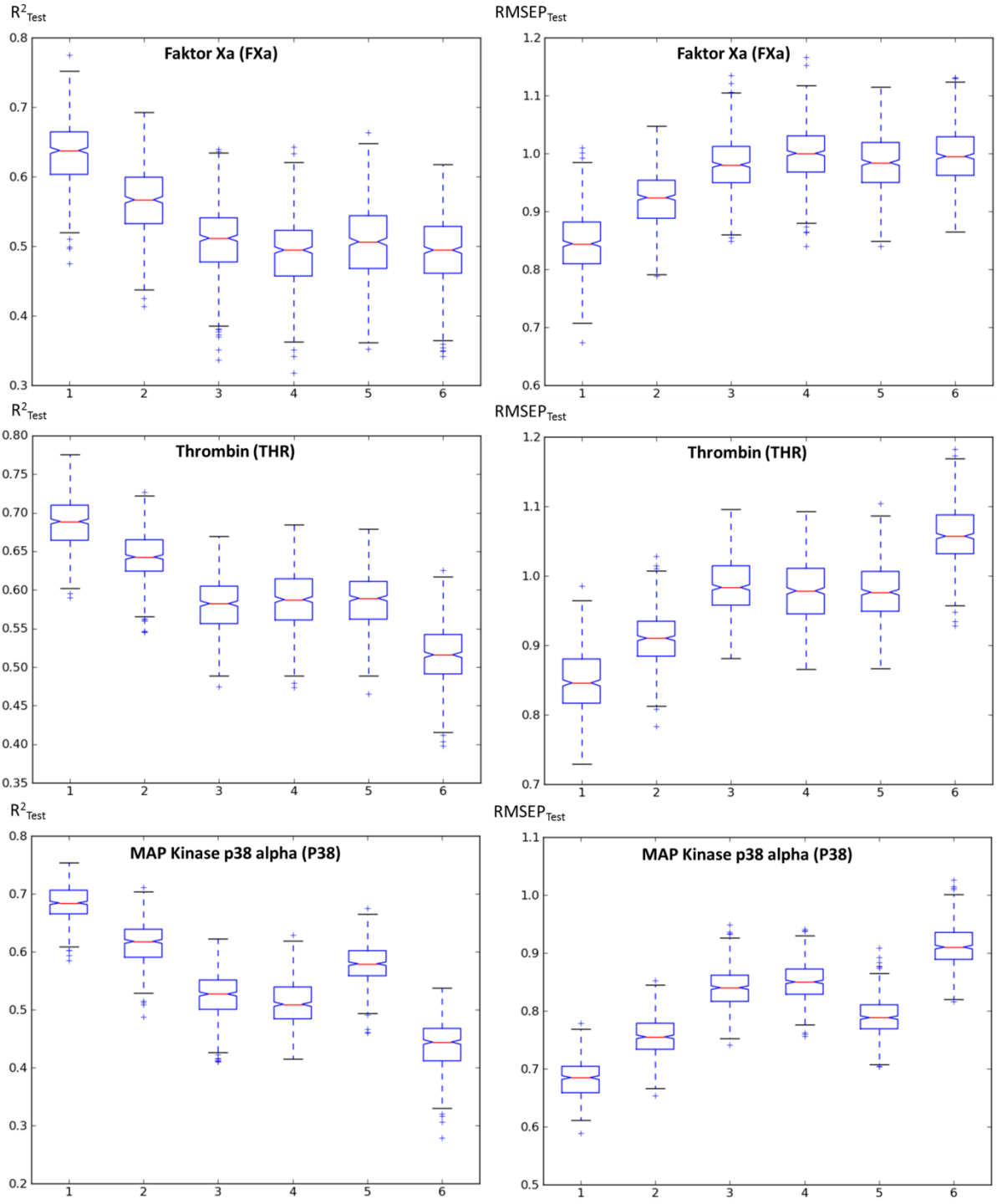


Abbildung 19.1. Boxplots der Ergebnisse (R^2_{Test} und $RMSEP_{Test}$) für die 500 Zufalls-Test-Stichproben der *k*NN-basierten Regression (k=3) (Datensätze: FXa, THR, P38). Details siehe Text. Legende: (1) inSARa, (2) ECFP4, (3) MACCSF, (4) MACCS, (5) FP2, (6) CATS2D.

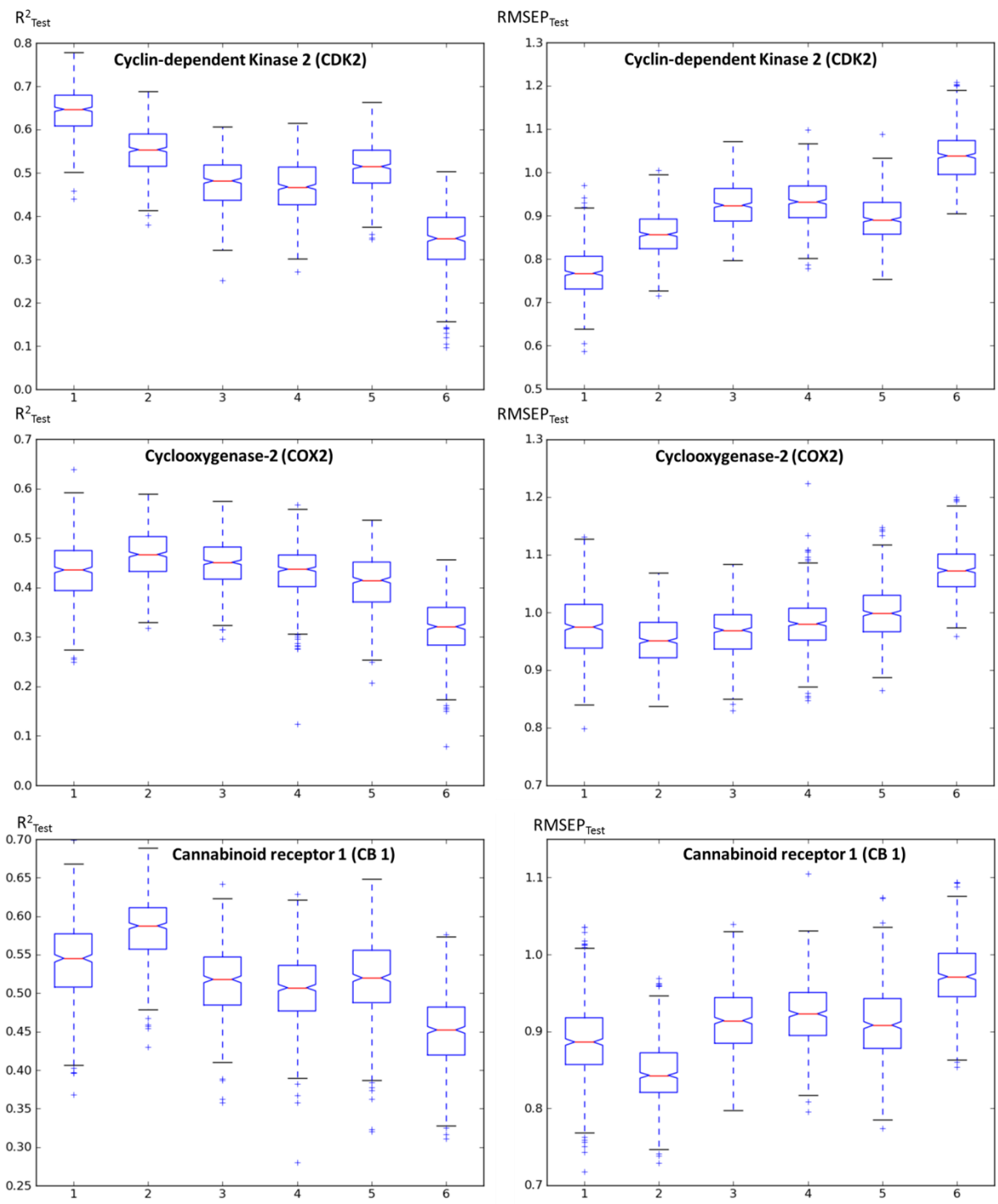


Abbildung 19.2. Boxplots der Ergebnisse (R^2_{Test} und $RMSEP_{Test}$) für die 500 Zufalls-Test-Stichproben der *k*NN-basierten Regression (k=3) (Datensätze: CDK2, COX2, CB1). Details siehe Text. Legende: (1) inSARa, (2) ECFP4, (3) MACCSF, (4) MACCS, (5) FP2, (6) CATS2D.

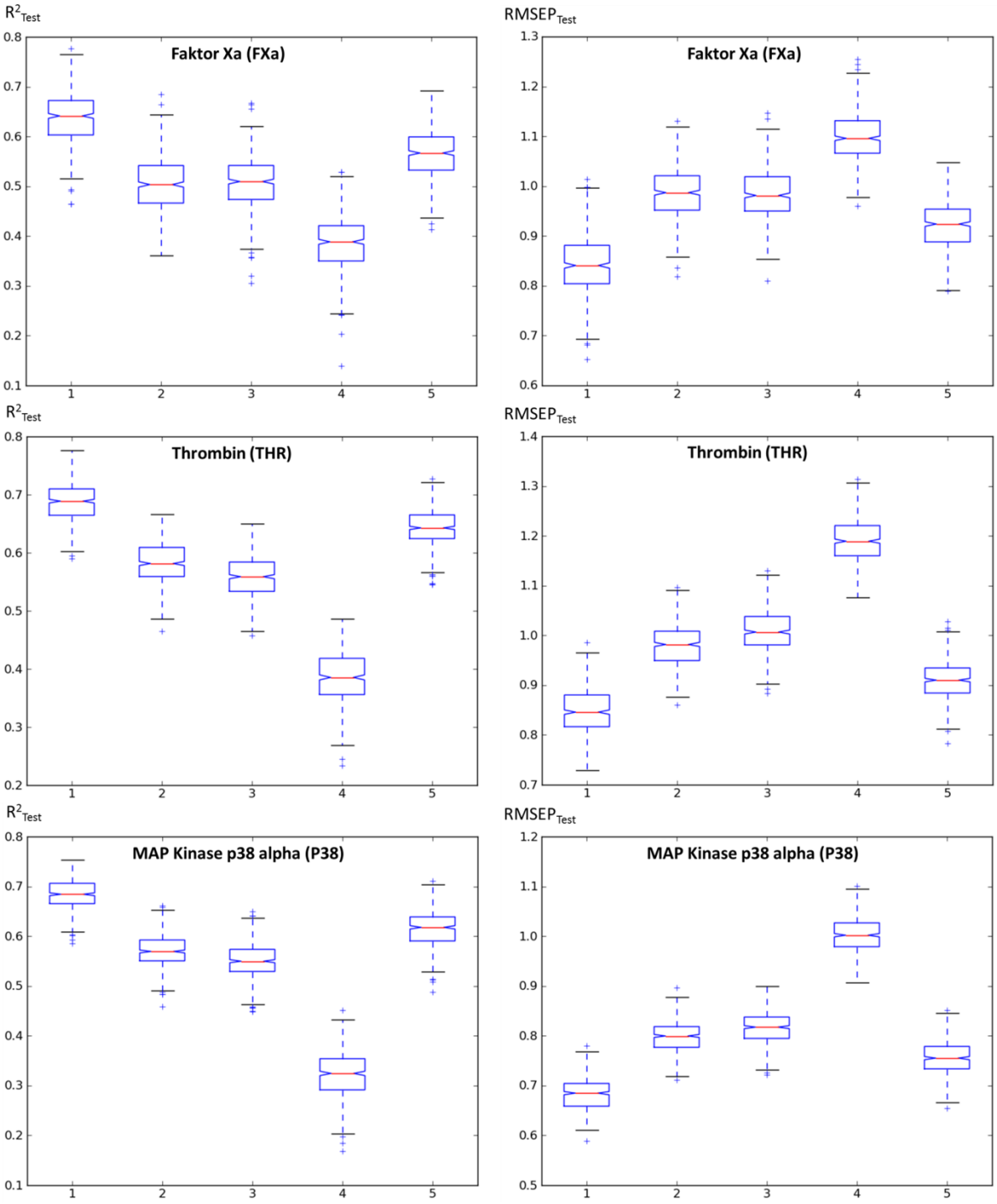


Abbildung 19.3. Boxplots der Ergebnisse (R^2_{Test} und $RMSEP_{Test}$) für die 500 Zufalls-Test-Stichproben der *k*NN-basierten Regression (k=3) (Datensätze: FXa, THR, P38). Details siehe Text. Legende: (1) inSARa, (2) RG_atompairs, (3) RG_atompairs_fuzzy, (4) RG_atom_count, (5) ECFP4.

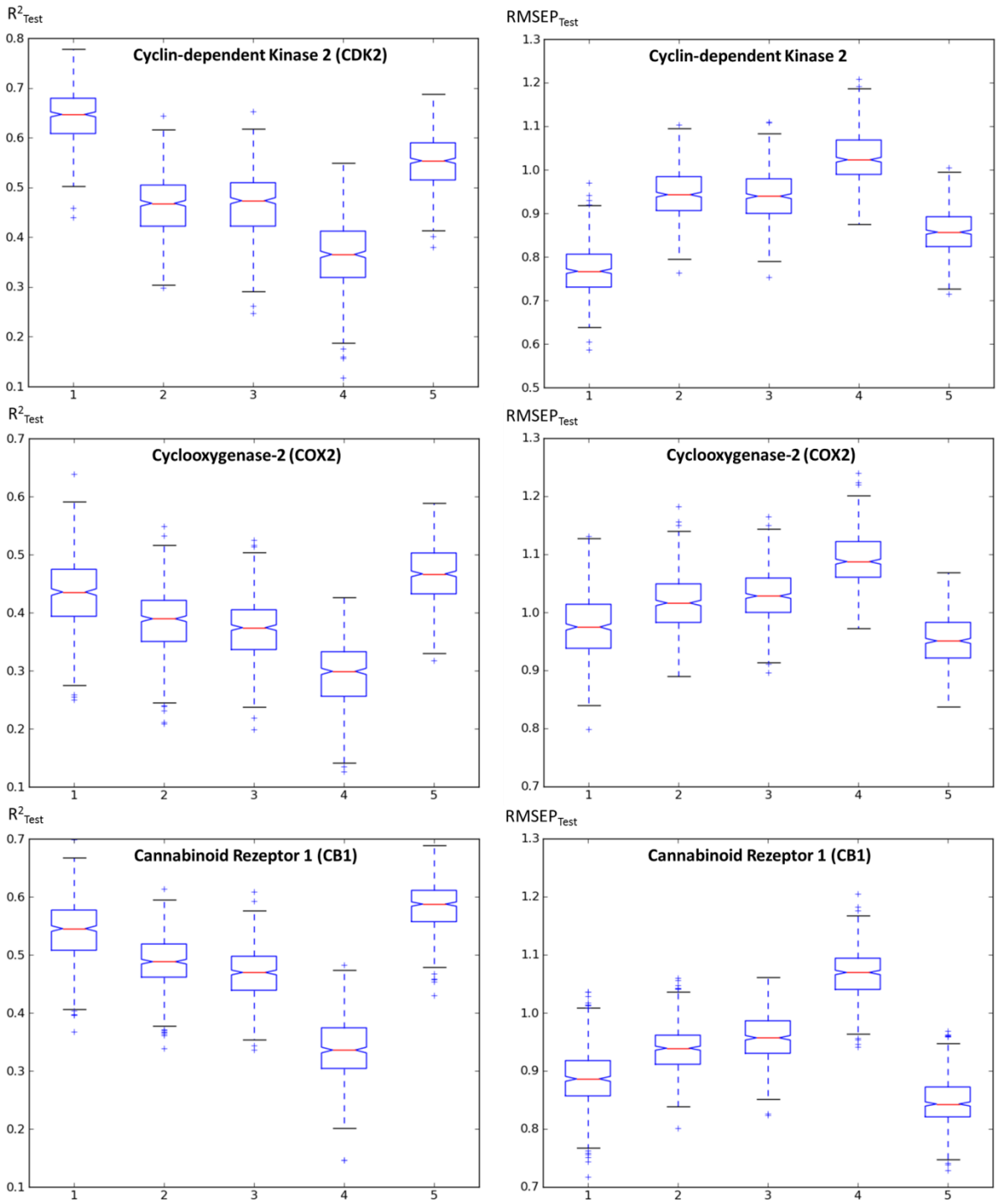


Abbildung 19.4. Boxplots der Ergebnisse (R^2_{Test} und $RMSEP_{Test}$) für die 500 Zufalls-Test-Samples der *k*NN-basierten Regression (k=3) (Datensätze: CDK2, COX2, CB1). Details siehe Text. Legende: (1) inSARa, (2) RG_atompairs, (3) RG_atompairs_fuzzy, (4) RG_atom_count, (5) ECFP4.

Das primäre Ziel dieser Analyse war es zu untersuchen, ob durch die andersartige Codierung der Moleküle und Erfassung der Ähnlichkeit bei inSARa eine verbesserte Auswahl der verfügbaren nächsten Nachbarn im chemischen Raum stattfindet, was sich in einer verbesserten Vorhersagegüte beim Verfahren des *k*NN äußern sollte. Bei dieser Analyse stand nicht die Evaluierung der absoluten Leistungsfähigkeit der inSARa-Methode, sondern der relative Vergleich mit verschiedenen Fingerprints im Vordergrund.

In Abbildung 19.1 bis Abbildung 19.4 ist erwartungsgemäß eine große Abhängigkeit der Ergebnisse von dem verwendeten Datensatz bzw. Target (Variabilität zwischen den Targets), Datensatzsplits (Variabilität innerhalb einer Methode für ein Target) und eingesetzten Methode (Variabilität zwischen den Methoden innerhalb eines Targets) zu erkennen. Wie auch im VS oft beschrieben, ist die Leistungsfähigkeit einer Methode stark Zielstruktur- oder Datensatz-abhängig, sodass unter den analysierten Verfahren keine Methode identifiziert werden kann, die den übrigen in jedem Fall bezüglich molekularer Repräsentation und Ähnlichkeitserfassung überlegen ist.

Für vier der untersuchten Targets (FXa, THR, CDK2, P38) kann eine deutliche Überlegenheit von inSARa auf Basis des relativen Fehlers (R^2_{Test}) und folglich auch auf Basis des absoluten Fehlers ($RMSEP_{Test}$) gegenüber den verwendeten Fingerprints festgestellt werden. Für die weiteren zwei Targets (COX2, CB1) sind die inSARa-Ergebnisse vergleichbar mit denen der Fingerprints, wobei der ECFP4 die besten Ergebnisse liefert. Die Ergebnisse für den topologischen Pharmakophor-Fingerprint CATS2D sind in allen Fällen (mit Ausnahme von FXa, wo die Ergebnisse vergleichbar mit MACCS, MACCSF und FP2 sind) deutlich schlechter als für die anderen Fingerprints. Der ECFP4-Fingerprint liefert von allen Fingerprints durchweg die besten Ergebnisse. Die Ergebnisse des MACCS- und MACCSF-Fingerprints sind zumeist vergleichbar. Bei einigen Targets kann eine marginale Überlegenheit der MACCSF beobachtet werden. Die Ergebnisse des FP2 sind am stärksten Target-abhängig. Für die Kinasen (P38, CDK2) sind die Ergebnisse mit dem ECFP4, für die restlichen Targets mit den MACCS Fingerprints vergleichbar.

Vergleicht man die verschiedenen Targets miteinander, so lässt sich feststellen, dass die besten Ergebnisse für die Kinase-Datensätze (P38 und CDK2), danach die Protease-Datensätze (Thrombin und FXa), gefolgt von dem CB1-Datensatz erzielt werden. Am schlechtesten schneidet der COX2-Datensatz ab.

Betrachtet man die absoluten Ergebnisse für inSARa so lässt sich feststellen, dass der beste RMSEP für P38 erhalten wird (Median etwa 0.68), gefolgt von CDK2 (Median etwa 0.77). Für FXa und Thrombin ist der RMSEP vergleichbar (Median etwa 0.84). Für CB1 liegt der Median etwa bei 0.88, während bei COX2 der Median der 500 Datensatz-Splits etwa bei 0.97 liegt. Der Median des R^2 variiert von 0.44 (für COX2) bis 0.69 (für Thrombin und P38).

Vergleicht man die für inSARa erzielten Ergebnisse mit denen verschiedener RG-Fingerprints (Abbildung 19.3 und Abbildung 19.4), so ist inSARa bei allen Targets auf Basis des absoluten und folglich auch auf Basis des relativen Fehlers den RG-Fingerprints überlegen. Der RG-Fingerprint „RG_atom_count“ ist bei allen Zielstrukturen den übrigen Verfahren deutlich unterlegen (Median des $R^2_{Test} < 0.4$, Median des $RMSEP_{Test}$ zwischen 1.0 und 1.2). „RG_atompairs“ und „RG_atomspairs_fuzzy“ liefern vergleichbare Ergebnisse und zeigen in allen Fällen etwas schlechtere Leistungsfähigkeit als der ECFP4-Fingerprint. Gegenüber dem topologischen Pharmakophor-Fingerprint CATS2D kann in allen Fällen eine deutliche Überlegenheit der beiden Atompaar-basierten RG-Fingerprints festgestellt werden.

Es ist anzumerken, dass in Einzelfällen Moleküle mit inSARa nicht vorgesagt werden konnten (keine MCSs vorhanden, die eine Substruktur des Molekül-RGs darstellen). Dies war jedoch nur durchschnittlich ein Molekül pro Testdatensatz-Stichprobe.

19.2. Diskussion

In Abschnitt 18.2 konnte gezeigt werden, dass eine gewisse Korrelation zwischen der Fingerprint- (v.a. ECFP4) und der RG-MCS-basierten basierten Ähnlichkeit besteht, jedoch auch Unterschiede festzustellen sind. Die Frage, ob diese Unterschiede in der Ähnlichkeits-Erfassung als positiv oder negativ für die SAR-Analyse zu bewerten sind, sollte mit dieser Analyse untersucht werden. Die Ergebnisse von FXa, THR, P38 und CDK2 zeigen, dass durch die andersartige Erfassung von Ähnlichkeit inSARa in der Lage ist, eine verbesserte Auswahl nächster Nachbarn zu treffen (geringerer Fehler). Auch bei den anderen beiden Targets ist inSARa mit den anderen Fingerprints vergleichbar und dem ECFP4-FP nur geringfügig unterlegen. Dies zeigt, dass für die inSARa-basierte SAR-Analyse Vorteile (d.h. sinnvollere Zusammengruppierung von „ähnlichen“ Molekülen) gegenüber den Fingerprints zu erwarten sind.

Der Vergleich mit den RG-Fingerprints belegt die Wichtigkeit der Berücksichtigung von Konnektivitäten bei der Erfassung von molekularer Ähnlichkeit. Der „RG_atom_count“ Fingerprint vergleicht Moleküle nur auf Basis des Vorhandenseins gleicher RG-Atomtypen, jedoch ohne Berücksichtigung jeglicher Konnektivität. Dies hat sich bei der Analyse in einer deutlichen Unterlegenheit gegenüber den übrigen Verfahren geäußert, wo Konnektivität zumindest partiell codiert wird. Die anderen beiden RG-Fingerprints erfassen im Vergleich zu der MCS-basierten Ähnlichkeit nur partiell molekulare Konnektivität auf Basis von RG-Atompaaren und die sie trennenden Distanzen. Der ECFP4-Fingerprint erfasst von allen untersuchten Fingerprints am besten lokale Ähnlichkeiten bzw. codiert molekulare Konnektivitäten auf Basis der Codierung der zirkulären Umgebung jeden Atoms. Dies erklärt, warum dieser Fingerprint die beste Leistung bei dieser Analyse zeigt.

Mit inSARa können einige Moleküle nicht vorhergesagt werden, sofern kein passender MCS, der eine Substruktur des RGs des Trainingsmoleküls darstellt, vorhanden ist. Für diese nicht-vorhersagbaren Moleküle existieren auf Basis der RG-MCSs keine ausreichend ähnlichen Trainingsmoleküle. Da sie außerhalb des Arbeitsbereiches des Vorhersagemodelles liegen, ist oftmals auch keine zuverlässige Vorhersage möglich. Eine Beschränkung der Vorhersage durch inSARa ist somit sinnvoll.

In Kapitel 2.2.2 wurde beschrieben, dass für heterogene IC_{50}-Daten ein mittlerer Fehler von mindestens 0,55 pIC_{50}-Einheiten und eine Standardabweichung von 0,68 pIC_{50}-Einheiten zu erwarten ist. Zudem sind die Ergebnisse des „RG_atom_count“ Fingerprints, die den Leistungsbereich einer schlechten Codierung andeuten, bei der Beurteilung der absoluten und relativen Leistungsfähigkeit der übrigen Verfahren (v.a. inSARa) zu berücksichtigen. In Anbetracht dieser Dimensionen und unter Berücksichtigung, dass die Vorhersagen auf keinerlei Lernprozess oder traininiertem Modell beruhen, liegt der absolute Fehler für die inSARa-Vorhersagen global betrachtet in einem akzeptablen Bereich. Um eine vergleichbare Leistungsfähigkeit wie die von trainierten QSAR-Modellen bzw. Methoden des überwachten maschinellen Lernens (z.B. SVM oder Random Forests) erwarten zu können, müssten beim

Aufbau der inSARa-Trainings-Netzwerke zusätzlich Elemente des überwachten Lernens eingebaut werden, die das vorhandene Wissen über die Bioaktivität bei Auswahl und Anordnung der Knoten im Netzwerk berücksichtigen würden. Eine weitere Möglichkeit die Vorhersagegüte zu verbessern, wäre die Definition einer Mindest-Ähnlichkeitsschwelle für die Berücksichtigung von nächsten Nachbarn. Eine Abschätzung der Zuverlässigkeit von Vorhersagen ist über die Berücksichtigung der Variabilität der zugehörigen Moleküle für einen MCS-Knoten bzw. bei FP-basierten Ansätzen die Variabilität der nächsten Nachbarn möglich.

Die Ergebnisse, die für einzelne Targets mit inSARa erzielt werden, sind ebenfalls stark von der Wahl der RG-Definition abhängig. Eine potentielle Ursache für schlechtere Ergebnisse im Vergleich zum ECFP4-FP bei einigen Targets könnte eine suboptimale Molekül-Codierung sein (eventuell zu hoher Abstraktionslevel). Wie schon mehrfach betont, sollte diese ggf. für das Target von Interesse jeweils unter Berücksichtung unter Umständen. vorhandener Besonderheiten (wie z.B. zusätzliche Codierung zinkbindender Gruppen bei Metalloenzymen, vgl. Abschnitt 6.5) optimiert werden. Eine Möglichkeit zu überprüfen, ob eine veränderte Codierung zur Verbesserung oder Verschlechterung führt, stellt eine solche Form der Analyse dar. Hiermit kann anhand objektiver Maßzahlen eine Verbesserung der Codierung bzw. die Güte eines resultierenden Netzwerkes beurteilt werden.

20. Ergebnisse und Diskussion: SAR-Interpretation

20.1. Analyse großer Datensätze aus der BindingDB am Beispiel von FXa

Im Folgenden wird die SAR-Analyse großer Datensätze mittels inSARa-Netzwerk beispielhaft gezeigt und diskutiert. Hierfür wurde als Musterbeispiel ein Datensatz aus der BindingDB ausgewählt, der nach der Vorbereitung aus 1736 Inhibitoren des Koagulations-Faktors Xa (Abk. FXa) besteht.

FXa nimmt eine wichtige Schlüsselrolle in der Blutgerinnungskaskade ein, da diese Serinprotease die Schnittstelle zwischen intrinsischem und extrinsischem Aktivierungsweg darstellt. Inhibitoren dieses Enzyms können therapeutisch somit als Antikoagulanzien verwendet werden. Wichtige Indikationen sind z.B. die postoperative Thromboembolie-Prophylaxe, Apoplex-Prophylaxe, Embolie-Prophylaxe bei Vorhofflimmern oder aber die Therapie von tiefen Venenthrombosen und Lungenembolien.[420–421] FXa stellt daher ein in den letzten Jahrzehnten intensivst beforschtes Target dar.[422] Rivaroxaban (Xarelto®) und Apixaban (Eliquis®) sind das Ergebnis dieser langjährigen Forschung. Diese ersten zugelassenen, oral verfügbaren FXa-Inhibitoren sind ein Beispiel für erfolgreiches rationales Wirkstoffdesign. Da die SARs von FXa-Inhibitoren hierdurch sehr gut erforscht sind und nicht nur zahlreiche Röntgen-Kristallstrukturen von Protein-Ligand-Komplexen verfügbar, sondern auch eine große Anzahl Bioaktivitätsdaten veröffentlicht sind, erschien FXa ein gutes Beispiel für die Validierung der SAR-Analyse von großen Datenmengen mittels inSARa.

Analyse des inSARa-Netzwerkes

Die Abbildung 20.1 und Abbildung 20.2 zeigen das resultierende inSARa-Netzwerk für den FXa-Datensatz unter Verwendung einer Mindest-MCS-Größe von 5 Pseudoatomen. Das gesamte inSARa-Netzwerk besteht aus 16 zusammenhängenden Komponenten, 36 Wurzel-MCSs und 660 weiteren MCS-Knoten. Aufgrund des Abbruch-Kriteriums von 2% nicht-repräsentierten Molekülen, sind 33 Moleküle nicht im Netzwerk abgebildet.

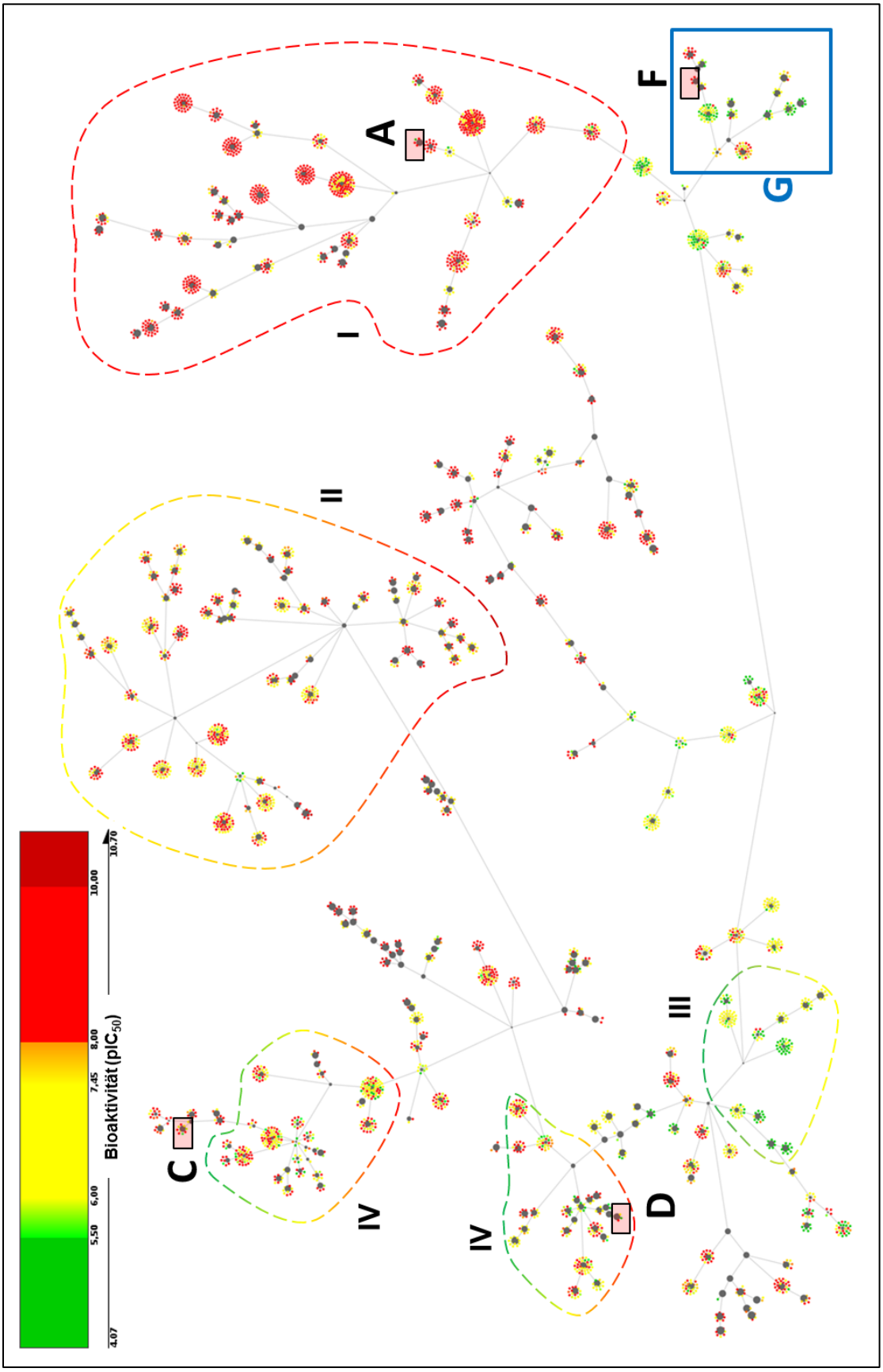


Abbildung 20.1. Größte Zusammenhangskomponente des inSARa-Netzwerkes des FXa-Datensatzes (pIC_{50}) aus der BindingDB (Parameter: Mindest-MCS-Größe = 5 RG Pseudoatome, Ausschlussliste = aktiv, Abbruchkriterium: ≤ 2% nicht-repräsentierte Moleküle). Das Layout wurde nach der automatischen Erstellung zur besseren Übersichtlichkeit manuell nachbearbeitet.

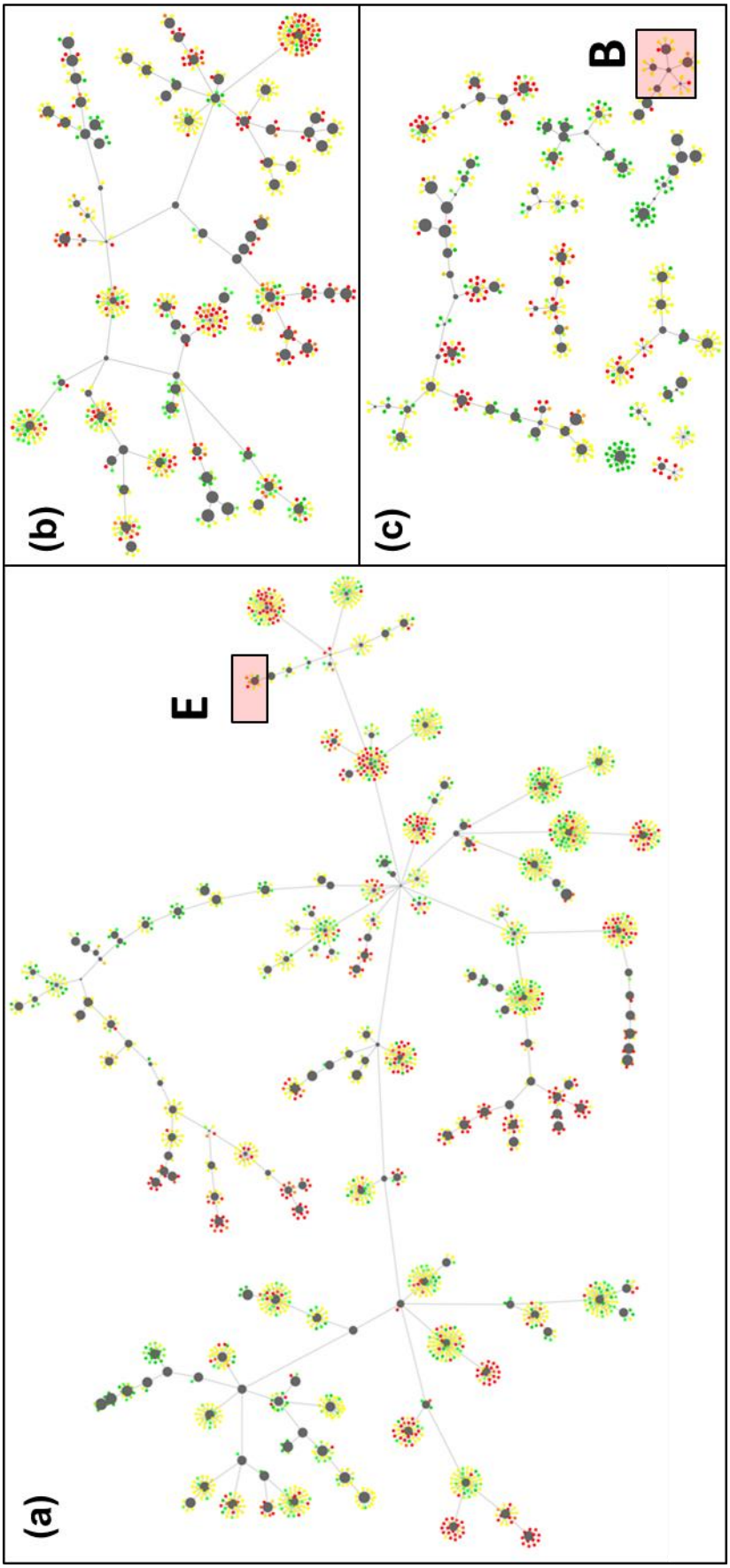


Abbildung 20.2. Die weiteren Komponenten des inSARa-Netzwerkes des FXa-Datensatzes (pIC_{50}) aus der BindingDB. (a) Zweitgrößte, (b) drittgrößte und (c) die übrigen 13 Komponenten.

20.1.1. Qualitative Analyse von SAR (Dis-)Kontinuität

Es lassen sich drei große Zusammenhangskomponenten im Netzwerk erkennen, durch die die Mehrheit der Datensatzmoleküle repräsentiert werden kann (vgl. Abbildung 20.1 und Abbildung 20.2a/b). Die größte dieser Komponenten ist in Abbildung 20.1 dargestellt. Hier lassen sich grob unterschiedliche Bereiche charakterisieren:

- *Homogenere Bereiche:*

  Sehr häufig kommen hier Netzwerk-Äste vor, die Moleküle mit ähnlicher Bioaktivität repräsentieren: Subnetzwerk I enthält beispielsweise hauptsächlich hochaktive Moleküle (die meisten terminalen Knoten sind rot gefärbt). In Subnetzwerk II finden sich v.a. mittel- und hochaktive Moleküle (die meisten terminalen Knoten sind gelb oder (dunkel) rot gefärbt). Bereich III hingegen ist geprägt von schwach- und mittelaktiven Molekülen (grün und gelbe Molekülknoten).

- *Heterogenere Bereiche:*

  Es lassen sich jedoch auch Netzwerk-Bereiche finden, die durch hohe Varianz in der Bioaktivität charakterisiert sind (vgl. die Subnetzwerke IV).

Global betrachtet zeigt FXa eine *heterogene Aktivitäts-Landschaft*. Dies ist konsistent mit dem berechneten globalen SAR-Index von 0,5 für diesen Datensatz (vgl. Tabelle 26.2 im Anhang). Im Netzwerk kann ein hoher Anteil an *kontinuierlichen SAR Bereichen* gefunden werden, in denen ähnliche Moleküle ähnliche Bioaktivitäten aufweisen und wo graduelle strukturelle Veränderungen nur zu moderaten Veränderungen in der Bioaktivität führen. Diese Feststellung ist ebenfalls konsistent mit dem hohen berechneten globalen SARI-Kontinuitäts-Wert (0,78). Es lassen sich jedoch auch viele MCS-Knoten mit hohen Bioaktivitäts-Unterschieden trotz ähnlicher molekularer Eigenschaften feststellen. Diese beobachtete *SAR Diskontinuität* ist ebenfalls konsistent mit dem hohen berechneten globalen SARI-Diskontinuitäts-Wert (0,88).

Es lässt sich des Weiteren beobachten, dass sich die größeren MCS-Knoten, insbesondere die terminalen MCS-Knoten, in den meisten Fällen weniger heterogen bezüglich der Bioaktivität verhalten als die kleineren MCS-Knoten oder die Wurzel-Knoten. Dies ist zu erwarten, da die Ähnlichkeit der Moleküle an MCS-Knoten, die kleinere MCSs repräsentieren, in der Regel geringer ist als die Ähnlichkeit der Moleküle an größeren bzw. terminalen MCS-Knoten.

20.1.2. Interaktive SAR-Analyse

Wendet man die in Abschnitt 16.1 vorgeschlagenen Regeln zur interaktiven SAR-Analyse an, so lassen sich verschiedene SAR-Informationen im Netzwerk identifizieren. Dies soll im Detail an den folgenden Beispielen (A bis G) veranschaulicht werden. Die Beispiele stammen jeweils aus den entsprechend markierten Bereichen in Abbildung 20.1 und Abbildung 20.2.

Es ist anzumerken, dass in manchen Abbildungen zusätzlich Bindetaschen-Informationen zu finden sind. Diese wurden nachträglich zur besseren Illustration eingefügt. Sie sind kein Bestandteil der standardmäßigen Visualisierung mittels Cytoscape. Zur Diskussion der einzelnen Beispiele wird die Standard-Nomenklatur nach SCHECHTER und BERGER zur Beschreibung der Bindetaschen bei Proteasen verwendet[423]. Hierbei werden die Seitenketten eines Peptidsubstrates mit P1, P2, P3,… (d.h. Pi) und die zugehörigen Bindetaschen der Protease als S1, S2, S3,… (d.h. Si) bezeichnet, wobei i zum N- bzw. i‘ zum C-Terminus des Substrates zunehmen. Vom N-Terminus aus gesehen befindet sich P1 direkt vor der zu spaltenden Peptidbindung, P1‘ direkt dahinter. Die Bindetasche der Protease für die Seitenkette der Aminosäure P1 wird demzufolge S1, die für die Seitenkette P4 entsprechend S4 genannt.

a) Lokale SAR-Analyse

1.) Sprunghafte SARs: Erkennen Interaktions-entscheidender Merkmale

Ein repräsentatives Beispiel für *sprunghafte SARs* (ein grünes Molekül umgeben von roten Knoten an einem einzelnen MCS-Knoten) ist in Abbildung 20.3 dargestellt. Der Austausch der Amidin-Gruppe (P1-Element, vgl. PDB-Code: 1EZQ; FXa) in Molekül 2 durch ein Amino-Isochinolin in Molekül 1 führt zu einem großen Bioaktivitätsverlust, obwohl der strukturelle Unterschied nur gering ist. Vergleicht man die (nicht in der Abbildung gezeigten) RGs dieser beiden Moleküle kann man sich diesen Aktivitätsabfall erklären. Das Amino-Isochinolin ist weniger basisch als das Benzamidin und somit weniger ionisiert. Da MOE die basische Funktionalität in Molekül 1 nicht protoniert und die SMARTS bei der RG-Umwandlung diese Amidin-Struktur aufgrund der Einbindung in den aromatischen Ring ebenfalls nicht als positiv ionisierbares Zentrum erkennen, enthält der RG von Molekül 1 im Gegensatz zu Molekül 2 keine PI-Eigenschaft. Da Ionisierung die ionische Wechselwirkung mit dem Asp_{189} in der S1-Tasche fördert, ist ein Absinken der Potenz zu erwarten. Anhand dieses Beispiels sieht man, dass sprunghafte SARs oftmals wertvolle Hinweise auf kritische molekulare Eigenschaften (z.B. positiv ionisierbare Gruppe) liefern können.

Betrachtet man sich die Moleküle 3 und 4 an, stellt man fest, dass es sich hier um analoge Strukturen zu Molekül 2 handelt. Diese unterscheiden sich nur durch verschiedene H-Brücken-Akzeptoren (Methylether, Methylester, Methylsulfon) am zum Benzamidin benachbarten Phenylring, die ohne Bioaktivitätsverlust gegeneinander ausgetauscht werden können. Diese drei Moleküle weisen aufgrund der gemeinsamen PI-Gruppe einen noch größeren gemeinsamen RG-MCS auf, weshalb sie an anderer Stelle im Netzwerk ebenfalls nochmal an einem reinen roten Knoten erscheinen. Diese sprunghafte SAR wird deshalb erkannt, weil Moleküle an mehreren Stellen im Netzwerk erscheinen können.

Ein Vorteil bei der Identifizierung von sprunghaften SARs mittels inSARa-Netzwerken im Vergleich zu Fingerprint-basierten Ansätzen ist, das man hier die ganze direkte chemische Nachbarschaft des Moleküls, das in eine sprunghafte SAR involviert ist, betrachten kann und nicht nur Molekülpaare verglichen werden. Man sieht im Fall des Beispiels A, dass an allen umliegenden Knoten in dem Subnetzwerk I zumeist hochaktive Moleküle gefunden werden. So lassen sich SAR Hotspots (sehr unebene Aktivitätslandschaft, d.h. hohe Variabilität der Bioaktivität auch in der Umgebung) und sprunghafte SARs (eine Klippe auf einer sonst eher

ebenen Fläche, d.h. ein Ausreißer bezüglich der Bioaktivität im Vergleich zur restlichen Umgebung) besser voneinander abgrenzen.

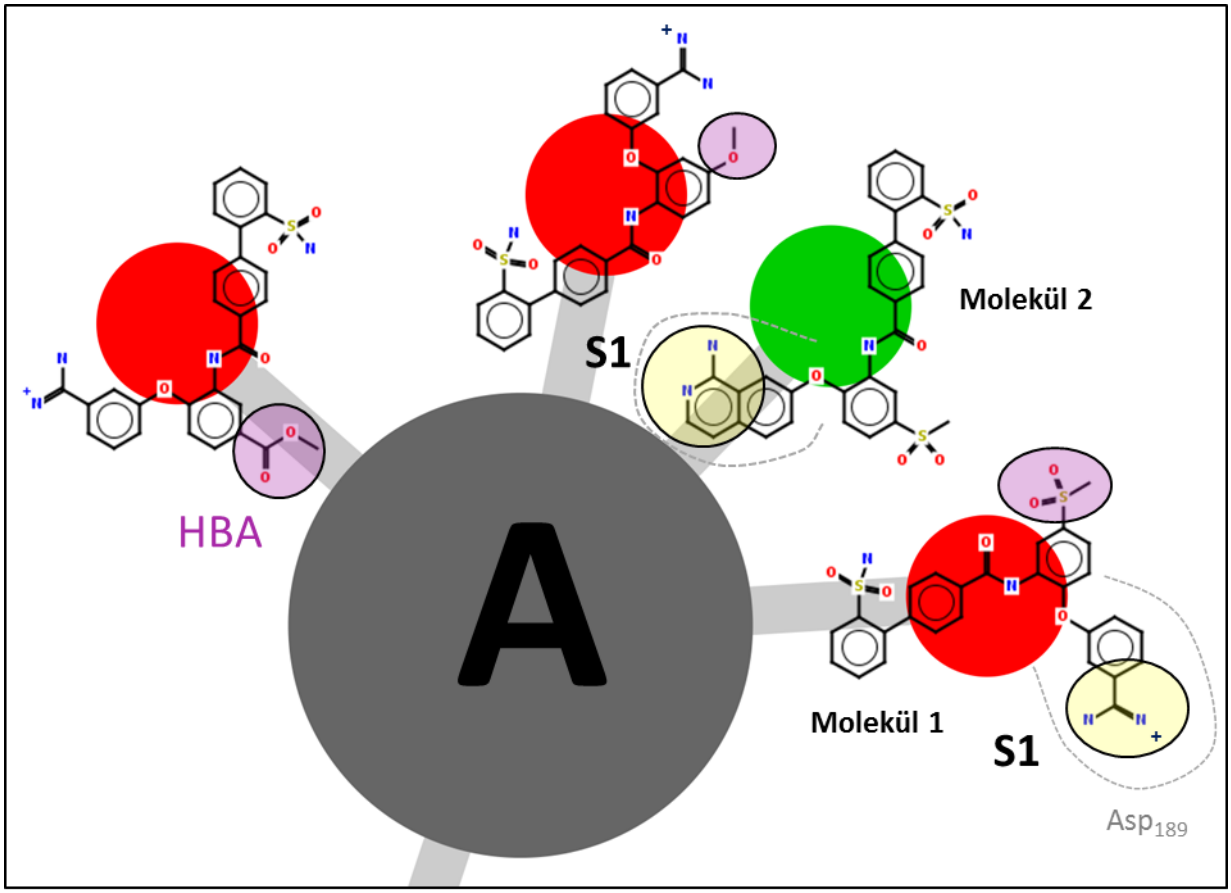


Abbildung 20.3. Identifikation von sprunghaften SARs in inSARa-Netzwerken. Die gelb markierten Gruppen unterscheiden Molekül 1 und 2, die die sprunghafte SARs repräsentieren. Die lila markierten funktionellen Gruppen stellen H-Brücken-Akzeptoren dar, die ohne Verlust an biologischer Aktivität ausgetauscht werden können.

2.) Bioisosterer Austausch: Optimierung weiterer Moleküleigenschaften

Hat man einen Treffer aus dem HTS in der „Hit-to-Lead“ Optimierung soweit optimiert, dass das Molekül eine bestimmte Mindest-Bioaktivität aufweist, ist es in der Regel notwendig weitere Parameter zu optimieren. Die Kenntnis von potentiell bioaktivitätserhaltenen, isofunktionellen Gruppen ist dabei von großem Wert (vgl. Abschnitt 2.5.2). Beispiele für die Identifizierung von *bioisosteren Gruppen* (gleiche Farben an einem ausreichend großen MCS-Knoten) sind in Abbildung 20.4 und Abbildung 20.5 dargestellt.

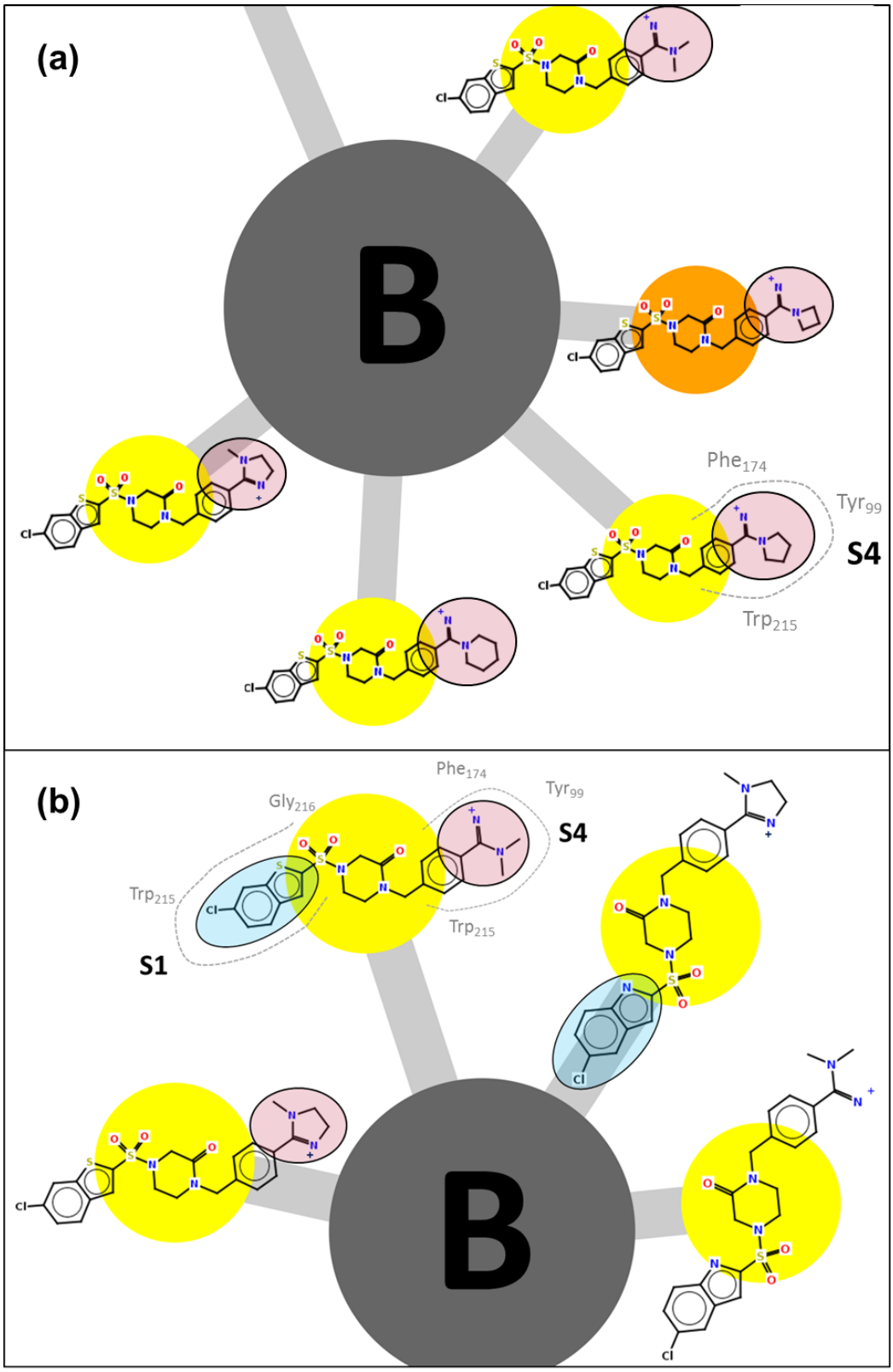


Abbildung 20.4. Identifizierung von bioisosterem Austausch in inSARa-Netzwerken. (a) Bioisostere positiv ionisierbare Gruppen (lila). Sie sind involviert in Kationen- π Interaktion in der S4-Tasche (vgl. PDB-Code: 2EI6; FXa). (b) Zwei bioisostere aromatische Ringsysteme (blau) und positiv ionisierbare Gruppen (lila). Die aromatischen Ringsysteme sind wichtige Eigenschaften für hydrophobe oder aromatische Interaktion mit der S1-Tasche (vgl. PDB-Codes: 2J34, 2EI6; FXa).

In Abbildung 20.4 sind an den beiden gezeigten terminalen MCS-Knoten alle Molekül-Knoten gelb gefärbt, d.h. alle Moleküle weisen mittlere Bioaktivität ($6 \leq pIC_{50} < 8$) auf. In der oberen Abbildung (a) sind verschiedene bioisostere positiv ionisierbare Gruppen markiert. Man sieht, dass die Amidin-Struktur unter Erhalt der biologischen Aktivität azyklisch oder in Ringsysteme verschiedener Größe eingebunden sein kann. In der unteren Abbildung (b) können zusätzlich noch zwei bioisostere aromatische Ringsysteme identifiziert werden. Auch bei diesen Beispielen sieht man, dass bestimmte Moleküle an mehreren Knoten erscheinen. Dies ermöglicht, dass je nach chemischer Nachbarschaft unterschiedliche SAR-Aspekte besser erfasst werden können.

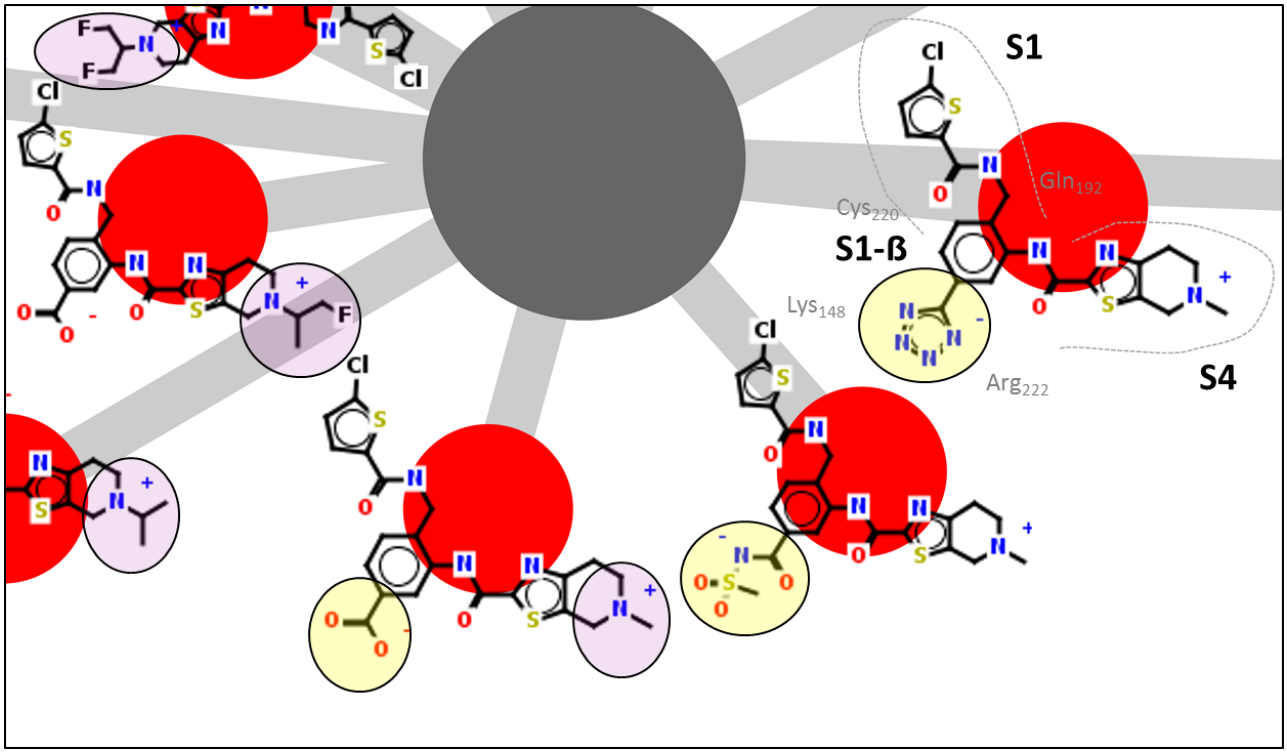


Abbildung 20.5. Identifizierung von bioisosterem Austausch in inSARa-Netzwerken. Bioisostere negativ ionisierbare Gruppen (gelb): Carbonsäure, Acylsulfonamid und Tetrazol. Sie gehen ionische Wechselwirkungen ein bzw. interagieren als HBA mit dem Lys148 und/oder Arg222 nahe der hydrophoben S1-ß-Tasche (sogenannte „Ester-bindende Tasche“[294, 289]) (vgl. PDB-Code: 3TK5; FXa). Die N-Methyl-Gruppe (lila) des Tetrahydropyridinothiazol-Ring (hydrophobe S4-Tasche) kann unter Erhalt der Bioaktivität gegen einen N-Isoproyl-Rest (lila) ausgetauscht werden. Die sterisch analogen Fluor-substituierten Isopropyl-Gruppen (lila) stellen bioisostere Gruppen dazu dar.

Anhand von Abbildung 20.5 soll ausführlicher die Bedeutung der Kenntnis von potentiell bioaktivitätserhaltenden Strukturmodifikationen oder bioisosteren Gruppen bei der Entwicklung neuer Arzneistoffstoffe veranschaulicht werden.

Das Einführen einer Carbonsäure-Gruppe am Phenylring (gelb markiert) führt zu hoch aktiven, oral verfügbaren zwitterionischen FXa-Inhibitoren, die eine sehr lange Wirkdauer aufweisen (bis zu 24h), was für die einmal tägliche Einnahme sehr vorteilhaft ist[424]. Ein weiteres Problem vieler FXa-Inhibitoren (vermutlich aufgrund der von lipophilen Bereichen umgebenen basischen Funktion) ist die hohe Affinität zum human Ether-à-go-go-Related

Gene (Abk. hERG) Kalium-Kanal[425]. Die starke hERG-Inhibition kann klinisch mit QT-Zeit-Verlängerung und lebensbedrohlichen Torsades-de-pointes-Arrhythmien einhergehen und ist ein wichtiger Grund für den Abbruch der Entwicklung neuer Substanzen[426]. Neben dem hERG-Kanal stellt CYP3A[427] ein weiteres wichtiges, sicherheitsrelevantes off-Target für die Prüfung neuer Arzneistoffe dar. Es konnte gezeigt werden, dass durch die Einführung einer negativ geladenen Carboxylgruppe aufgrund elektrostatischer Wechselwirkungen nicht nur die Affinität zum hERG-Kanal verloren geht[425], sondern auch die CYP3A4-Inhibition deutlich reduziert ist[428]. Die Einführung einer Carboxyl-Gruppe geht jedoch mit verringerter Lipophilie und Permeabilität einher. Bioisosterer Austausch kann diese Eigenschaften verbessern unter Erhalt der genannten pharmakologischen Aktivität[424]. In Abbildung 20.5 weisen alle Moleküle an dem MCS-Knoten hohe biologische Aktivität auf (rote Molekül-Knoten, $pIC_{50} \geq 8$). Man erkennt, dass sowohl der Tetrazolring als auch die Acylsulfonamid-Gruppe zur Carbonsäure bioisostere Gruppen darstellen.

Des Weiteren ist in Abbildung 20.5 zu erkennen, dass die N-Methyl-Gruppe am protonierten basischen Amins, sowohl gegen eine Isopropyl-Gruppe (höhere metabolische Stabilität[424]) als auch gegen dessen fluorierte Derivate (lila Markierung) unter Erhalt der Bioaktivität ausgetauscht werden können. Die fluorierten Isopropyl-Derivate haben sehr ähnliche sterische und lipophile Eigenschaften wie eine Isopropyl-Gruppe, jedoch ermöglichen sie es aufgrund der hohen Elektronegativität der Fluor-Atome die Basizität des Amins zu reduzieren, was zwar zu verringerter Löslichkeit, aber auch zu gesteigerter Permeabilität und somit oraler Bioverfügbarkeit führt[424].

3.) SAR Hotspots zur Leitstruktur-Optimierung

Die Abbildung 20.6 und Abbildung 20.7 stellt jeweils ein Beispiel für *SAR Hotspots* (Auftreten von verschiedenen Farben an einem MCS-Knoten) in inSARa-Netzwerken dar.

Aus Abbildung 20.6 lassen sich molekulare Elemente, die entscheidend die Bioaktivität beeinflussen, abgeleitet werden. Für S4-Optimierung ist die Ringgröße wichtig, damit die Moleküle optimal die S4-Tasche ausfüllen (Nr. 1). Man kann sehen, dass sterisch anspruchsvollere Ringe zu einer sterischen Hinderung führen (Nr. 2-4)[429], was in einer starken Reduktion der Bioaktivität resultiert. Für S1-Optimierung scheinen eine zusätzliche OH-Gruppe (Nr. 7) oder ein Phenylring (Nr. 6) vorteilhaft aufgrund der zu beobachtenden Affinitätszunahme im Vergleich zu dem mittelaktiven Molekül (Nr. 9), wo diese funktionellen Gruppen jeweils fehlen.

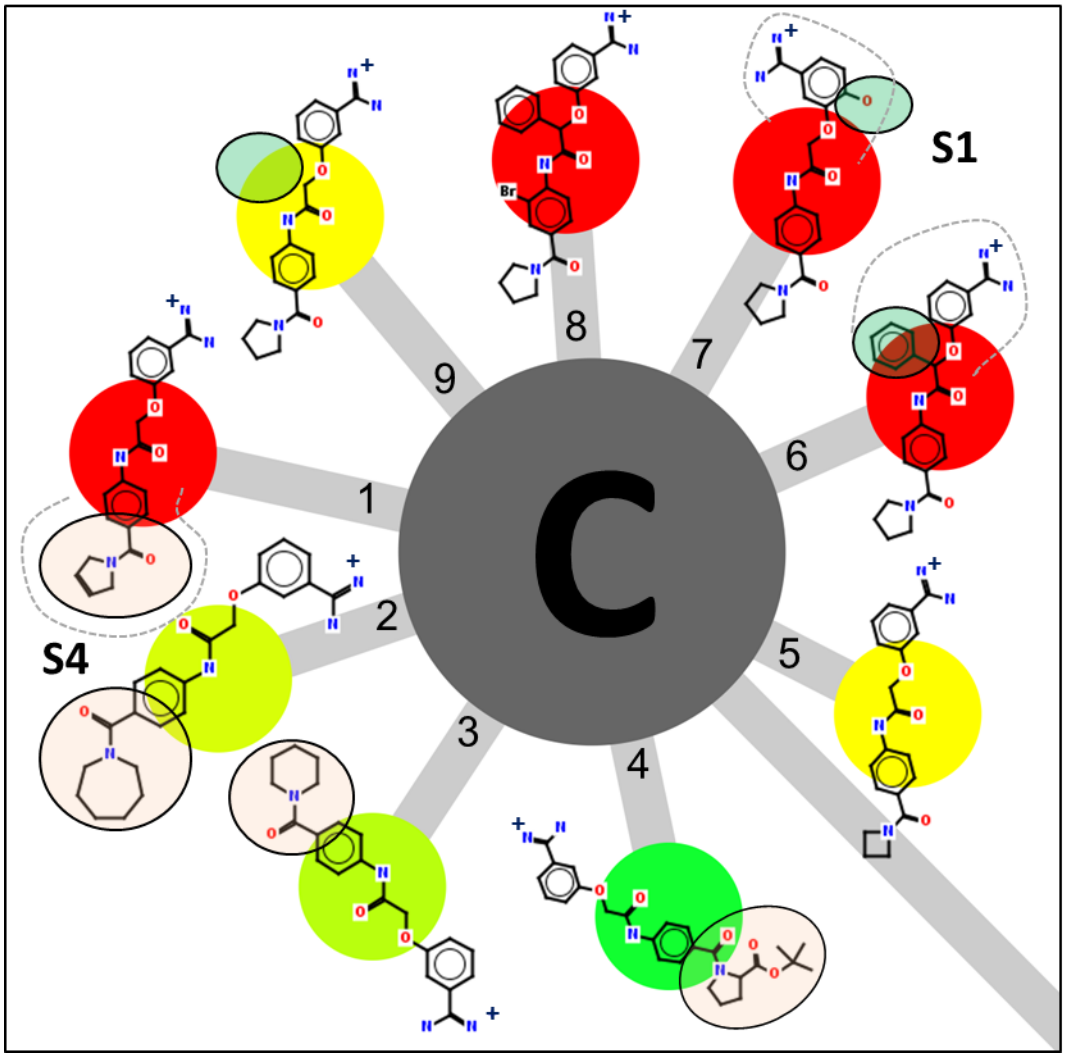


Abbildung 20.6. Identifizierung eines SAR Hotspots in inSARa-Netzwerken. Die Größe des Ringes, der mit der S4-Tasche der FXa-Bindestelle interagiert, ist entscheidend für die Bioaktivität. Voluminöse Ringe (Nr. 2-4) führen zu sterischer Hinderung, wodurch die unteren Moleküle nur schwach aktiv sind. Das Einfügen einer Hydroxyl-Gruppe in para-Position zum Amidin (Nr. 7) oder eines aromatischen Ringes am Linker zwischen den beiden Phenylringen (Nr. 6) scheinen potentiell günstige molekulare Optimierungen für eine verbesserte Interaktionen mit der S1-Tasche zu sein.

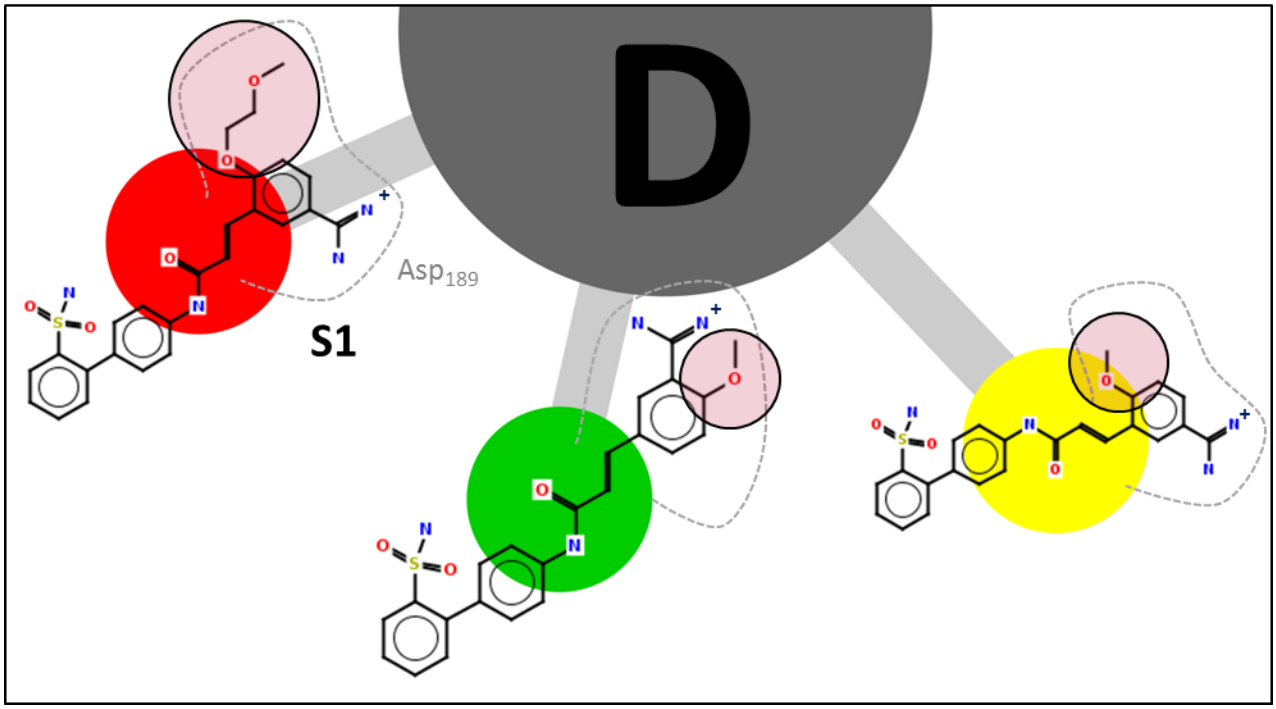


Abbildung 20.7. Ein weiteres Beispiel für einen SAR Hotspot. Der Substituent und seine Position am Benzamidin-Ring beeinflussen stark die biologische Aktivität (S1-Optimierung).

Abbildung 20.7 zeigt, dass die Substitution des Benzamidin-Ringes in der S1-Tasche stark die biologische Aktivität beeinflusst[430]. Substitution an Position 4 ist gegenüber Position 2 bevorzugt. Das größere 4-Methoxyethoxy-Analogon ist stärker potent als das Molekül mit 4-Methoxy-Substitution.

4.) Stereochemie-bedingte SARs und (Nicht-)Bioisosterie

Wie bei der RG-Umwandlung beschrieben (vgl. Abschnitt 10.2.3) wird Stereochemie nicht im RG codiert. Daher findet man Stereoisomere an demselben MCS-Knoten im inSARa-Netzwerk. Dies ist auch in Abbildung 20.8 zu erkennen, wo der Einfluss von Stereochemie und nicht-bioisosterem Austausch auf die Bioaktivität veranschaulicht ist. Alle Moleküle werden durch den gleichen RG codiert, der zugleich auch den MCS darstellt. Dennoch ist eine hohe Varianz bezüglich der Bioaktivität zu beobachten. Es lässt sich ableiten, dass das R-Enantiomer jeweils das Eutomer darzustellen scheint und stärkere Interaktion mit dem Enzym eingeht. Analysen führen dies auf die für die Rezeptor-Interaktion energetisch günstigere äquatoriale Stellung des Aminomethyl-Restes (R-Konfiguration) am Pyrrolidin-Ring im Vergleich zu axialen Stellung (S-Konfiguration) zurück[431].

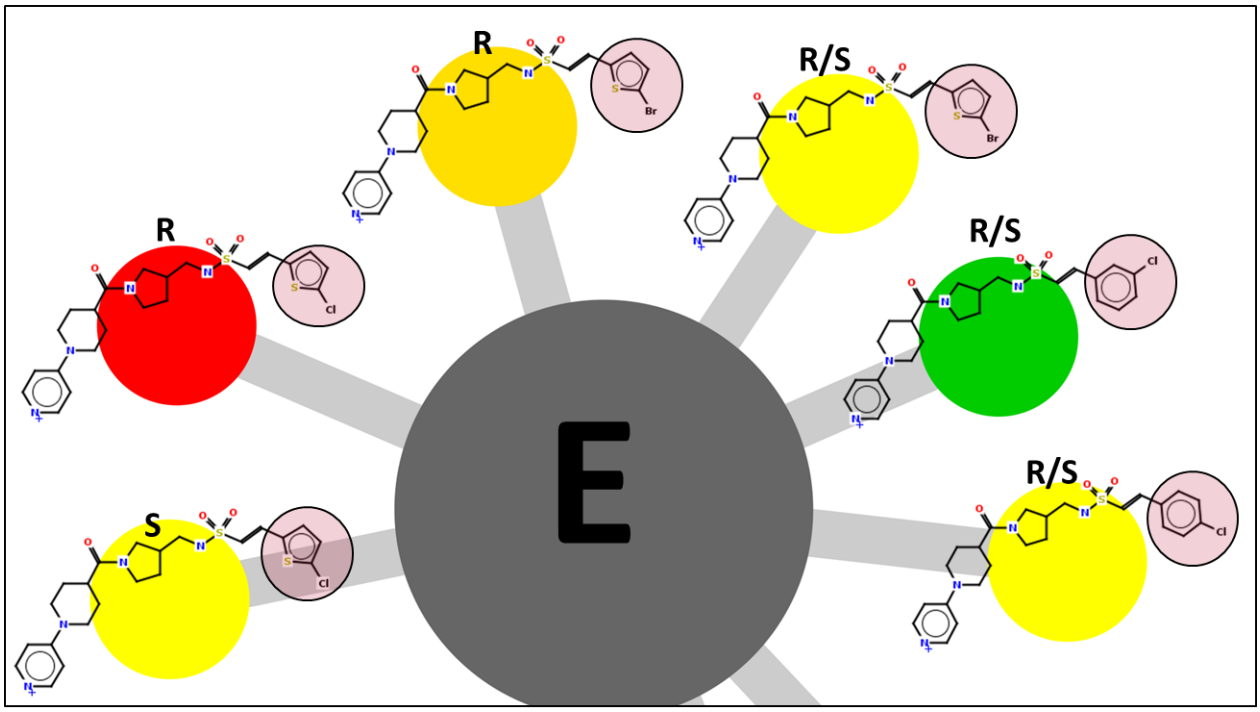


Abbildung 20.8. Einfluss von (nicht in RGs codierter) Stereochemie auf die biologische Aktivität. R-Konfiguration am chiralen Zentrum des Pyrrolidin-Ringes wird bevorzugt. Bei den markierten Ringsystemen (lila) handelt es sich um einen nicht-bioisosteren Austausch.

5.) Zielstrukturabhängigkeit von (Nicht-)Bioisosterie

In der Literatur wird Thiophen als typisches zum Phenylring bioisosteres Ringsystem beschrieben. In Abbildung 20.8 ist jedoch festzustellen, dass die Chlor-Phenyl-Analoga nicht bioisoster zu den Chlor- oder Brom-Thiophenyl-Molekülen sind.

Ähnliches zeigt auch der SAR Hotspot in Abbildung 20.9. Die Chlor-Phenyl-Analoga und das Chlor-Thiophenyl-Derivat stellen keinen bioisosteren Austausch dar. Ursache für dieses Verhalten ist, dass das Chlor-Atom an einer Bioaktivitäts-entscheidenden Schlüssel-Interaktion mit dem aromatischen Ring des Tyr-228 in der S1-Tasche beteiligt ist (vgl. Kristallstruktur von Rivaroxaban, PDB-Code: 2W26; FXa).[432]

Diese Beispiele bestätigen, dass Bioisosterie (wie oft beschrieben, z.B: BIRCHALL et al.[216], MEANWELL[154]) in der Regel Zielstruktur-abhängig analysiert werden muss und generelle Übertragungen auf andere Targets schwierig sind.

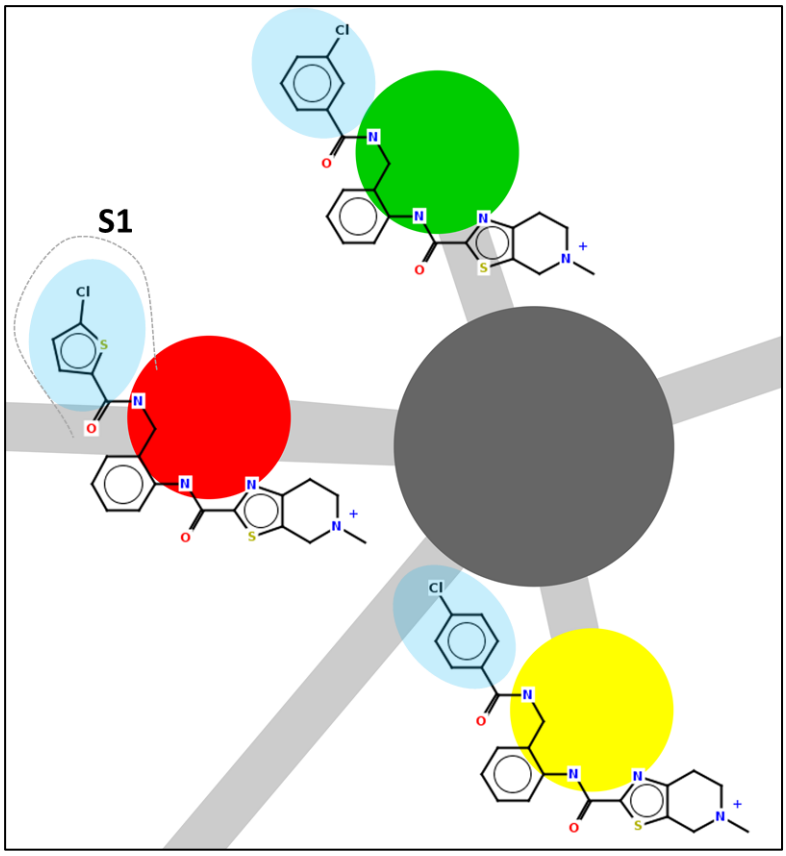


Abbildung 20.9. SAR Hotspot im FXa-inSARa-Netzwerk für die Identifizierung von nicht-bioisosteren Ringsystemen (blau markiert) in der S1-Tasche (vgl. PDB-Code: 2W26; FXa). Details siehe Text.

Abbildung 20.10 zeigt, dass selbst SAR Hotspots zur Identifizierung von bioisosteren Ringsystemen geeignet sein können. Alle lila markierten Ringsysteme können in der Leitstruktur-Optimierung bioisoster ausgetauscht werden. Verantwortlich für beobachteten Bioaktivitätsdifferenzen sind die Substituenten oder deren Position (blau markiert) am adjazenten Phenylring. Eine 3-Trifluormethylgruppe führt zu starker Affinitätssteigerung (IC_{50} = 1nM) im Vergleich zur 3-Fluorgruppe. Eine 2-Methylgruppe hingegen führt zu starker Affinitätsabnahme (IC_{50} = 1.26µM).

Halogenatome können wie in Kapitel 6.5 beschrieben in verschiedenartige, schwer abschätzbare Interaktionstypen involviert sein. In Leitstruktur-Optimierungs-Projekten werden (wie auch dieses Beispiel zeigt[432]) verschiedene Halogenatome, Halogen-haltige Gruppen sowie Methyl- oder längere Alkyl-Gruppen häufig gegeneinander ausgetauscht (Optimierung der physikochemischen Eigenschaften z.B. Hydrophobie oder aber Interaktionsoptimierung). Bei inSARa werden diese verschiedenen Gruppen als Zn-Pseudoatom codiert. Trotz der geringen Spezifität dieser Codierung (keine Berücksichtigung potentieller pharamakophorer Eigenschaften) hat sich dieses Codierungsschema in einer Vielzahl an SAR-Analysen als sinnvoll erwiesen. Falls jedoch die Rolle bestimmter Gruppen an dem zu analysierenden Target bekannt sein sollte, ist es jederzeit möglich diese Codierung durch Definition entsprechender SMARTS und weiterer Pseudoatome spezifischer zu gestalten.

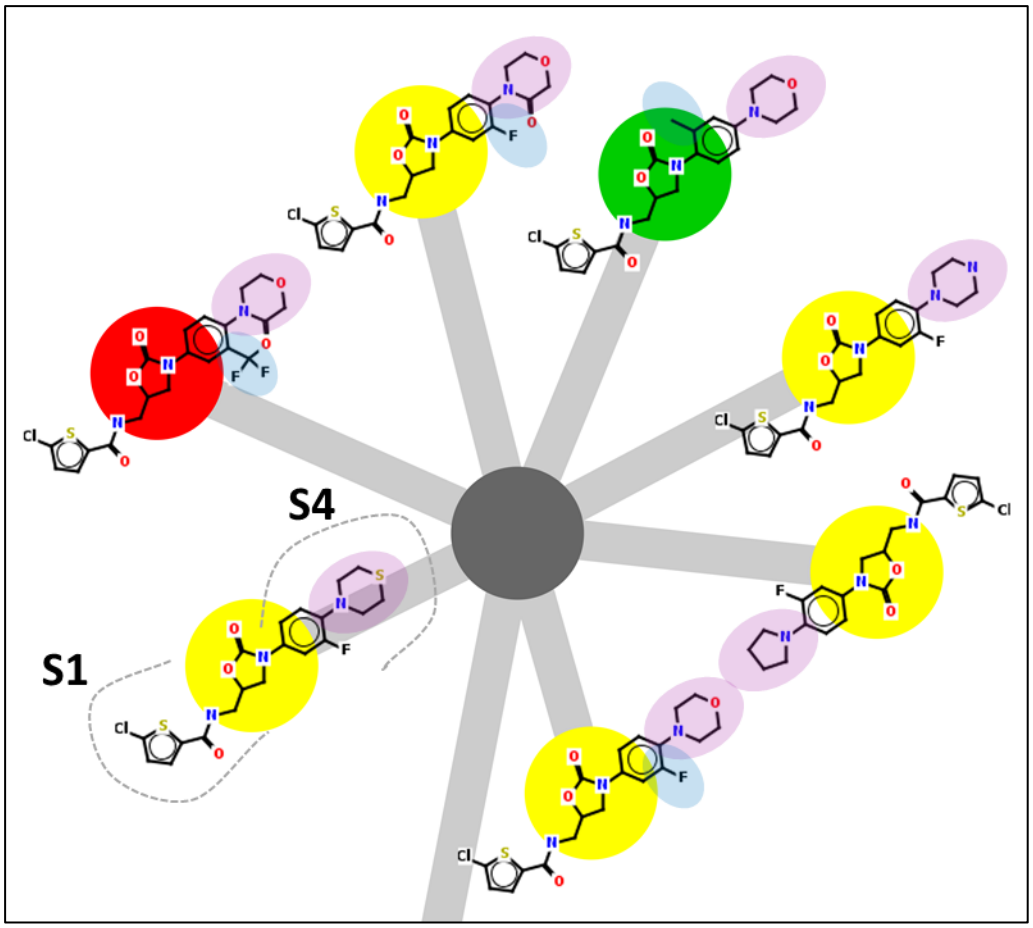


Abbildung 20.10. SAR Hotspot im FXa-inSARa-Netzwerk zur Identifizierung bioisosterer Ringsysteme in der S4-Tasche (vgl. PDB-Code: 2W26; FXa). Alle lila markierten Ringsysteme können bioisoster ausgetauscht werden. Für die Bioaktivität entscheidend ist der Substituent am benachbarten Phenylring (blau markiert). Details siehe Text.

6.) Hinweis auf potentielles Vorhandensein mehrerer Bindungsmodi

Der MCS-Knoten F in Abbildung 20.11 zeigt vier hochaktive FXa-Inhibitoren (rote Knoten). Molekül A weist eine Benzamidin-Gruppe, die (wie in Abbildung 20.3 gesehen) ein essentielles Merkmal für die Interaktion mit der S1-Tasche im aktiven Zentrum des Enzyms darstellt. Molekül B, C und D hingegen unterscheiden sich an dieser Stelle von Molekül A durch eine Chlor- oder Brom-Pyridinyl-Gruppe. Diese molekularen Eigenschaften gehen normalerweise nicht die bekannte Schlüssel-Interaktion mit dem Asp_{189} ein, sodass hier ein Hinweis auf das Vorhandensein eines weiteren Bindungsmodus in der S1-Tasche gegeben wird. Diese Hypothese lässt sich nicht nur mit Kristallstrukturen aus der PDB (z.B. 2P3T; FXa) belegen, sondern ist auch in der Literatur gut beschrieben[433].

Die Tanimoto-Ähnlichkeit für Molekül A und B sowohl unter Verwendung von MACCS Keys (Tc = 0,67) als auch des ECFP4-Fingerprints (Tc = 0,41) sind relativ niedrig. In Fingerprint-basierten Ähnlichkeits-Netzwerken (z.B. NSGs[193]) gibt es Kanten nur zwischen Knoten, die einen bestimmten Ähnlichkeits-Schwellenwert überschreiten. Unter Verwendung üblicher Schwellenwerte (MACCS Keys 0,75 und ECFP4 0,55) existieren keine Kanten zwischen den Knoten von Molekül A und B, sodass kein direkter Vergleich dieser beiden Moleküle bei der

visuellen Inspektion der Netzwerke erfolgt. Ein Rückschluss auf einen weiteren Bindungsmodus würde nur auf indirektem Weg erfolgen.

Das entsprechende Fingerprint-basierte Ähnlichkeits-Netzwerk (unter Verwendung des ECFP4-Fingerprints und eines Tc-Schwellenwertes von 0,55) ist in Abbildung 20.12 dargestellt. Man sieht, dass die Moleküle, die die beiden Bindungsmodi repräsentieren, deutlich voneinander entfernt sind. Im Gegensatz zu inSARa wird keine direkte Beziehung zwischen diesen Molekülen hergestellt. inSARa hingegen weist den Benutzer daraufhin das Vorhandensein verschiedener Bindungsmodi für diese in der Gesamtheit sehr ähnlichen Moleküle, die sich hauptsächlich in einem einzelnen strategisch wichtigen Merkmal unterscheiden, zu überprüfen.

Es ist anzumerken, dass die Moleküle B-D (Bindungsmodus 2) aufgrund des Vorhandenseins einer größeren gemeinsamen Substruktur an einem weiteren MCS-Knoten im inSARa-Netzwerk (ohne Molekül A) zusammengruppiert werden. Aufgrund der mehrfachen Repräsentation von Molekülen im Netzwerk an verschiedenen MCS-Knoten wird es jedoch ermöglicht diese drei Moleküle im Vergleich mit Molekül A zu betrachten und die oben beschriebenen Rückschlüsse zu ziehen.

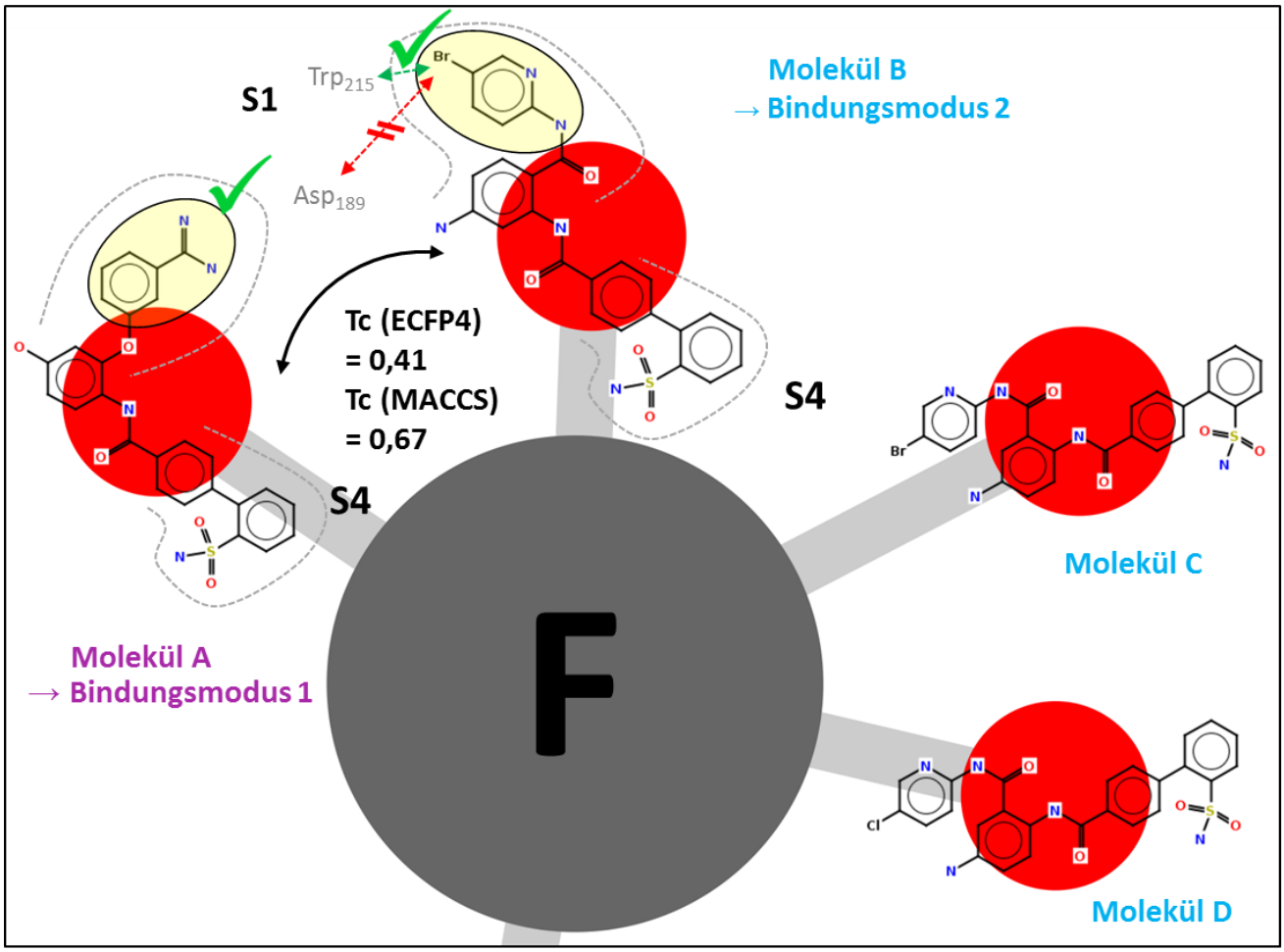


Abbildung 20.11. Rückschlüsse auf verschiedene Bindungsmodi in der S1-Tasche der Bindestelle von FXa durch das RG-MCS-basierte Zusammengruppieren von Molekül A und B an einem gemeinsamen inSARa-MCS-Knoten. Aufgrund der geringen Fingerprint-Ähnlichkeit würden Molekül A und B in Fingerprint-basierten Netzwerken keine direkten Nachbarn darstellen.

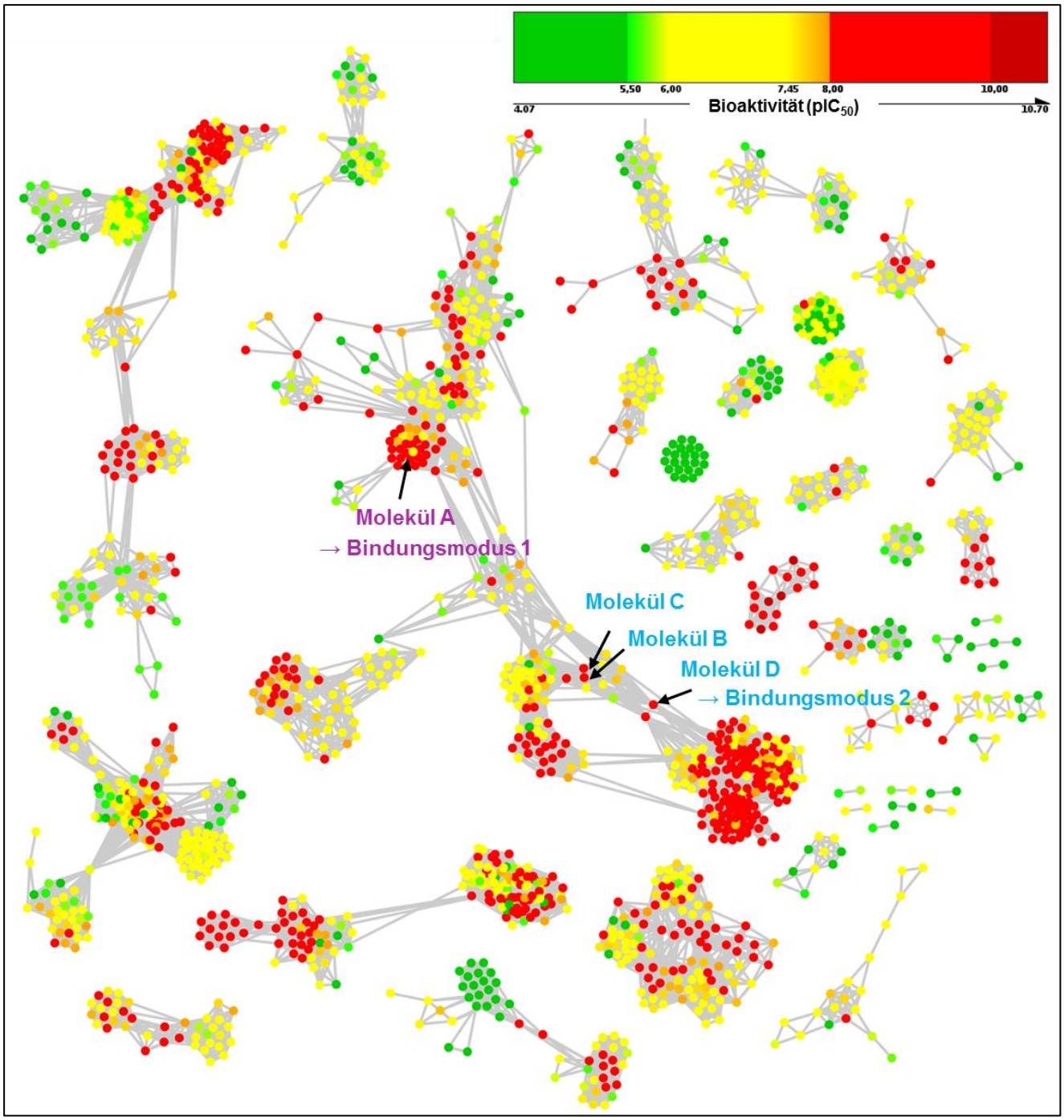


Abbildung 20.12. Fingerprint-basiertes Ähnlichkeits-Netzwerk für den FXa-Datensatz (Fingerprint = ECFP4, Tc ≥ 0,55 für das Erstellen von Kanten). Das Netzwerk besteht aus 43 Komponenten. Die Moleküle aus Abbildung 20.11, die verschiedene Bindungsmodi in der S1-Tasche repräsentieren, befinden sich gemeinsam in der größten Komponente. Im Gegensatz zum inSARa-Netzwerk wird jedoch keine direkte Beziehung hergestellt.

b) Subnetzwerk-Analyse: Activity Switches

Eine Besonderheit von inSARa-Netzwerken ist, dass sie nicht nur für die Identifizierung von SAR-Trends bezogen auf einzelne Moleküle ermöglichen, sondern auch das Erkennen von generellen SARs, z.B. in Form von sogenannten Activity Switches, erleichtern. Dies stellt einen deutlichen Unterschied zu Fingerprint-basierten Verfahren dar, die aufgrund des paarweisen Ähnlichkeitsvergleichs und der berechneten Ähnlichkeitswerte weniger zum Erkennen dieser Trends geeignet sind.

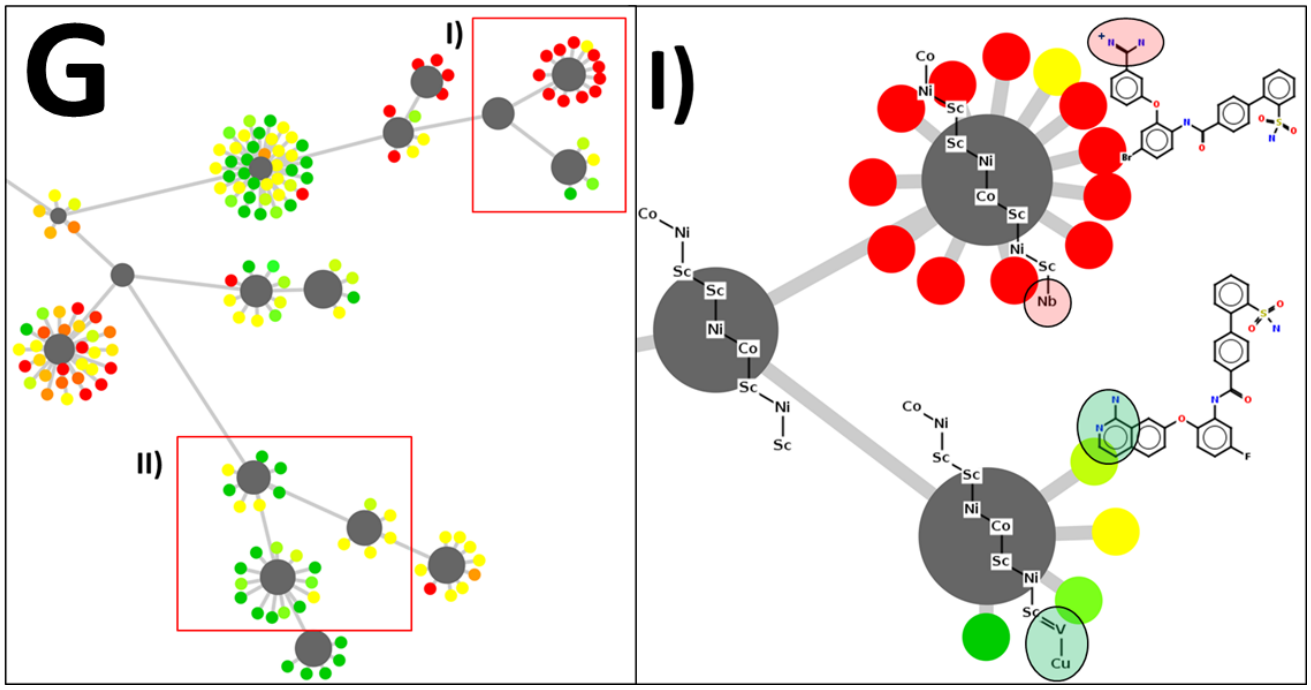


Abbildung 20.13. Identifizierung von Activity Switches in inSARa-Netzwerken. Der Netzwerkausschnitt links oben stammt aus dem blau markierten Bereich G in Abbildung 22.1. I) und II) stellen zwei Typen von Activity Switches dar. Das Beispiel I) (links) zeigt die Trennung von schwach und hoch aktiven Molekülen durch das Hinzufügen eines Merkmals (PI/nicht-PI-Eigenschaft). Die zugehörigen RG-MCSs sind auf dem MCS-Knoten abgebildet und einige Moleküle sind beispielhaft neben den entsprechenden Molekül-Knoten abgebildet. Die pharmakophoren Eigenschaften und entsprechenden molekularen Merkmale, die für die Bioaktivitäts-Auftrennung verantwortlich sind, sind markiert und werden exemplarisch gezeigt.

Die Identifizierung von Activity Switches mittels inSARa wird anhand der Beispiele I) und II) in Abbildung 20.13 bzw. Abbildung 20.14 veranschaulicht. Die beiden Beispiele stammen aus dem blau markierten Ausschnitt G der Hauptkomponente im FXa-Netzwerk (vgl. Abbildung 20.1). Auf der linken Seite von Abbildung 20.13 ist dieser Ausschnitt noch einmal im Detail abgebildet.

In Beispiel I) kann man sehen, dass die beiden größeren MCS-Knoten auf der rechten Seite hoch- und schwachaktive Moleküle klar voneinander trennen, obwohl nur ein Merkmal im Vergleich zum kleineren Vorgänger-MCS-Knoten verändert ist. Vergleicht man die (mittels chemViz an den entsprechenden MCS-Knoten abgebildeten) RG-MCSs, so lässt sich dieser SAR-Trend leicht interpretieren. Das Hinzufügen einer positiv ionisierbaren Gruppe (codiert

durch das Pseudoatom Nb im RG, im Molekül handelt es sich z.B. um eine Amidin-Gruppe) am Phenyl-Ring führt zu hochaktiven Molekülen. Nicht-PI-Gruppen hingegen führen zu schwachaktiven Molekülen. Dies ist analog zu dem Beispiel für sprunghafte SARs in Abbildung 20.3. Hier sieht man abermals, dass das mehrfache Auftreten von Molekülen im Netzwerk Vorteile hat, denn so können sich SAR-Trends auf verschiedene Weise zeigen. Wichtige Information wird so weniger leicht übersehen. Dies geht jedoch mit einer erhöhten Netzwerk-Komplexität einher. Bei noch größeren Datensätzen kann dies jedoch auch die Interpretation aufgrund von Informationsüberflutung behindern.

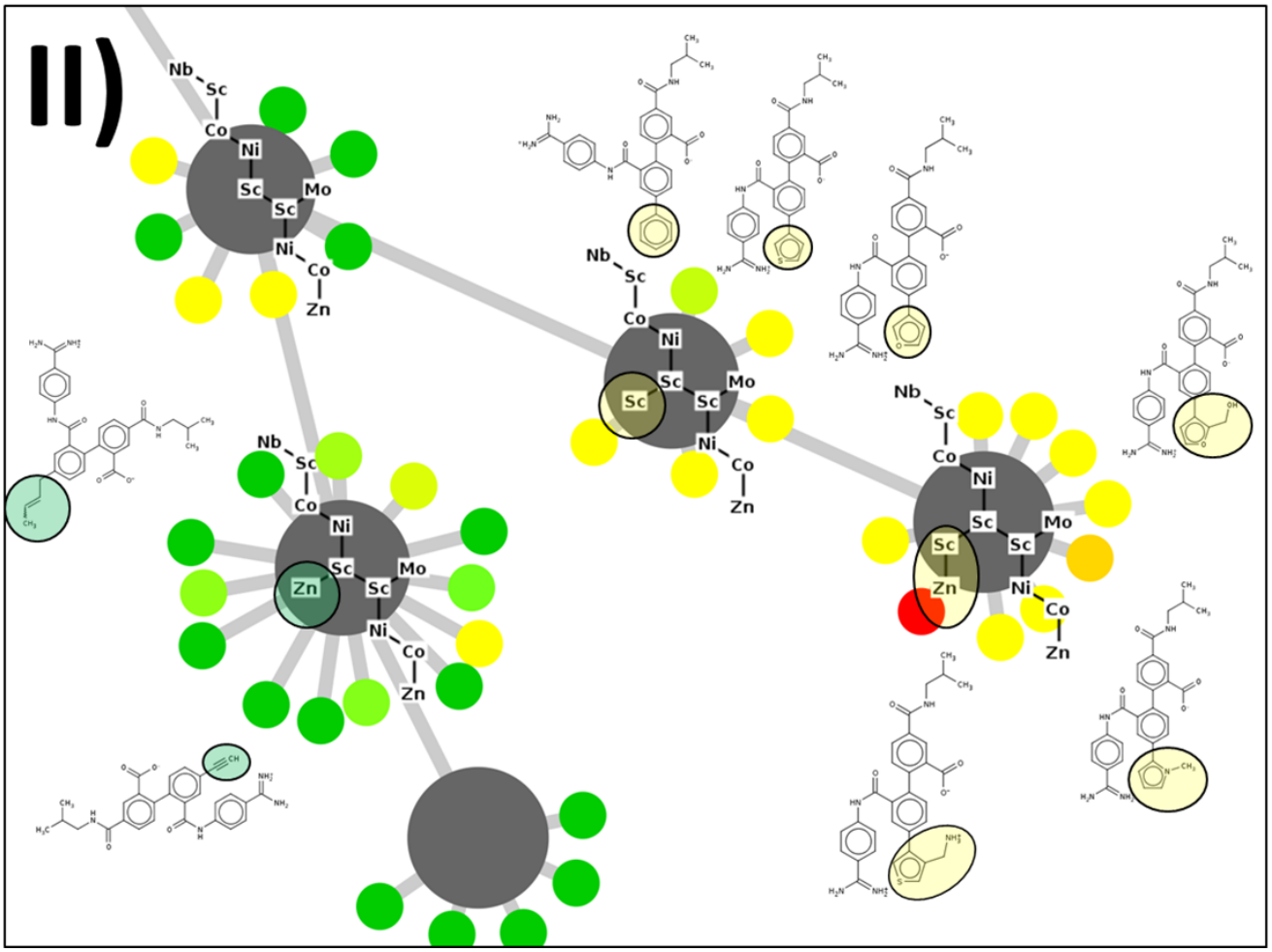


Abbildung 20.14. Weiteres Beispiel für einen Activity Switch (Ausschnitt II) aus Abbildung 20.13). Durch Anfügen einer aromatischen Gruppe an den Phenylring mittelaktive Moleküle werden von schwach aktiven Molekülen getrennt, die durch eine azyklische terminale Gruppe (z.B. Alkylgruppen) charakterisiert sind. Die zugehörigen RG-MCSs sind auf dem MCS-Knoten abgebildet und einige Moleküle sind beispielhaft neben den entsprechenden Molekül-Knoten abgebildet. Die pharmakophoren Eigenschaften und entsprechenden molekularen Merkmale, die für die Bioaktivitäts-Auftrennung verantwortlich sind, sind markiert und werden exemplarisch gezeigt.

Das zweite Beispiel II) zeigt einen weiteren Typ von Activity Switch. Während die Eigenschaften, die durch den kleinsten MCS-Knoten (oben in Abbildung 20.14) repräsentiert werden, nicht ausreichen, um schwach- und mittelaktive Moleküle zu trennen (gelbe und grüne Molekül-Knoten), trennen die größeren benachbarten Nachfolger-MCS-Knoten beide

Aktivitätsklassen klar. Wie im vorherigen Beispiel kann durch Betrachten der entsprechenden RG-MCSs die Interpretation erleichtert werden. Das Hinzufügen einer weiteren aromatischen Gruppe (codiert durch das Pseudoatom Sc) zu einem der aromatischen Ringe führt überwiegend zu Molekülen mit mittlerer Bioaktivität, während sich durch verschiedene azyklische terminale Gruppen, wie z.B. Alkyl- oder Alkenyl-Gruppen (codiert durch das Pseudoatom Zn) keine Optimierung der Bioaktivität erreichen lässt.

20.1.3. Vergleich von inSARa-Netzwerken und NSG-ähnlichen Netzwerken

Vergleicht man die beiden SAR-Netzwerk-Typen (inSARa in Abbildung 20.1 und Abbildung 20.2, das Fingerprint-basierte Ähnlichkeits-Netzwerk in Abbildung 20.12), so können einige Stärken und Schwächen dieser unterschiedlichen Ansätze herausgestellt werden. Beide Methoden ermöglichen systematische SAR-Visualisierung und -Analyse, sie erfassen jedoch molekulare Ähnlichkeit durch unterschiedliche Arten der molekularen Repräsentation und Ähnlichkeitsmaße. Ein Vorteil von inSARa ist die hierarchische, baumartige Netzwerk-Struktur, die die Netzwerk-Navigation und die intuitive SAR-Interpretation erleichtert. Diese klare Struktur fehlt in den FP-basierten Netzwerken. Mit der Hilfe der RG-MCS-Knoten kann (wie in den obigen Beispielen gezeigt) die angenommene chemische Ähnlichkeit zwischen verschiedenen Molekülen, die mit diesem einzelnen MCS-Knoten verbunden sind, in den inSARa-Netzwerken erklärt werden. In FP-basierten Netzwerken hingegen bleiben die Gründe für die Ähnlichkeit oftmals unklar. Dies limitiert die SAR-Analyse. Durch das Aufzeigen gemeinsamer pharmakophorer Eigenschaften in inSARa wird die Interpretation nicht nur erleichtert, sondern es können auch abstraktere Beziehungen, die nicht durch die gebräuchlichen Fingerprint-Typen wie ECFP4 oder MACCS Keys codiert werden, deutlich werden. Jedoch können Moleküle manchmal aufgrund des Abstraktionslevels auch fälschlicherweise zusammengruppiert werden. Besonders bei kleineren MCSs (im Vergleich zur RG-Größe) ist die Wahrscheinlichkeit dafür groß. Für die effektive SAR-Interpretation von FP-basierten Netzwerken sind die in Abschnitt 2.6.4 aufgezeigten Erweiterungen (z.B. SAR Pathways) notwendig. Ein Vorteil von FP-basierten Netzwerken ist die geringere Komplexität (aufgrund des einmaligen Vorkommens eines jeden Moleküls in dem resultierenden Netzwerk) im Vergleich zu inSARa, wo mehrfaches Auftreten erlaubt ist. Ein weiterer Vorteil ist der geringere Rechenaufwand im Vergleich zu inSARa, wo die Berechnungen aufgrund der mit den MCS-Bestimmungen einhergehenden Komplexität deutlich anspruchsvoller sind. Zusammenfassend lässt sich feststellen, dass sich beide Konzepte einander ergänzen, da beide Vor- und Nachteile aufweisen, die nicht im jeweils anderen Ansatz vorhanden sind. Eine maximale Ausbeute aus der SAR-Analyse kann daher erwartet werden durch die vernünftige Kombination der Informationen, die von beiden Netzwerk-Typen zur Verfügung gestellt werden.

20.2. Beispiel-Anwendung: weitere große Datensätze aus der BindingDB

Anhand der nachfolgenden Beispiele aus weiteren Datensätzen sollen weitere Besonderheiten von inSARa veranschaulicht werden und es soll gezeigt werden, dass die anhand des FXa-Netzwerkes gezeigte SAR-Interpretation auch auf andere Zielstrukturen übertragbar ist.

20.2.1. Beispiel CDK2

Die Serin-/Threonin-Kinase Cyclin-abhängige Kinase 2 (engl. cyclin-dependent kinase 2, Abk. CDK2) spielt wie alle Protein-Kinasen eine wichtige Rolle in einer Vielzahl von physiologischen Zellprozessen, insbesondere der Regulation des Zellzyklus.[434] Aufgrund ihrer essentiellen Rolle in Zellproliferation ist ihre Dysregulation ein wichtiger Faktor bei der Entstehung und Progression von Tumorerkrankungen.[435] Selektive CDK2-Inhibitoren stellen somit eine vielversprechende Therapieoption für maligne Tumore dar.[434, 436] CDK2 stellt daher wie viele andere Kinasen ein vielbeforschtes Target in den vergangenen Jahren dar, sodass sowohl eine Vielzahl von Bioaktivitätsdaten als auch Kristallstrukturen verfügbar sind.[436–438] Ein selektiver Inhibitor ist bisher jedoch trotz einiger Kandidaten in der klinischen Prüfung bisher nicht bis zur Zulassung gelangt.[436, 439–440]

Der CDK2-Datensatz (IC_{50}) aus der BindingDB besteht nach der Vorbereitung aus 1575 Inhibitor-Molekülen der CDK2. In Abbildung 20.15 ist das resultierende inSARa-Netzwerk gezeigt (Einstellungen: Mindest-MCS-Größe = 5 RG-Atome, Abbruch-Kriterium: 2% nicht-repräsentierte Moleküle, Ausschlussliste = aktiv). Es besteht aufgrund der erhöhten Mindest-MCS-Größe aus 76 zusammenhängenden Komponenten (zum Vergleich: bei Mindest-MCS-Größe von 3 RG-Atomen nur 3 zusammenhängende Komponenten), 105 Wurzel-Knoten und 676 weiteren MCS-Knoten. 31 Moleküle sind nicht im Netzwerk repräsentiert. Für die Erstellung des automatischen Layouts ist die größere Zahl Komponenten mit weniger Knoten von Vorteil. Die Analysen verschiedener Datensätze haben gezeigt, dass so leichter ästhetische Layouts (ohne manuelle Nachbearbeitung) erhalten werden.

SAR-Charakterisierung

Nach der SAR-Charakterisierung von PELTASON und BAJORATH[135] ergibt sich für den Datensatz ein sehr hoher globaler SARI-Kontinuitäts-Wert (0,89) und ein mittlerer globaler SARI-Diskontinuitäts-Wert (0,56). Daraus resultiert ein Gesamt-SAR-Index von 0,67, was typisch für *kontinuierlichere SARs* ist. Dieser Eindruck wird auch bei der qualitativen Analyse des inSARa-Netzwerkes bestätigt. Die meisten Komponenten sind sehr homogen gefärbt, was auf geringe Bioaktivitätsunterschiede auch bei leichten strukturellen Veränderungen hindeutet. Nur wenige Subnetzwerke sind heterogener Natur.

Ein ähnlicher Eindruck ergibt sich auch bei dem in Abbildung 20.16 dargestellten Fingerprint-basierten Ähnlichkeits-Netzwerk (ECFP4-FP, Schwellenwert von Tc ≥ 0,55). Wie auch im inSARa-Netzwerk sind v.a. die kleinen Komponenten sehr homogen, bei den größeren Komponenten sind ebenfalls einige Fälle zu beobachten, wo z.B. ein hochaktives mit einer Gruppe von schwachaktiven Molekülen verbunden ist oder umgekehrt (SAR-Diskontinuität).

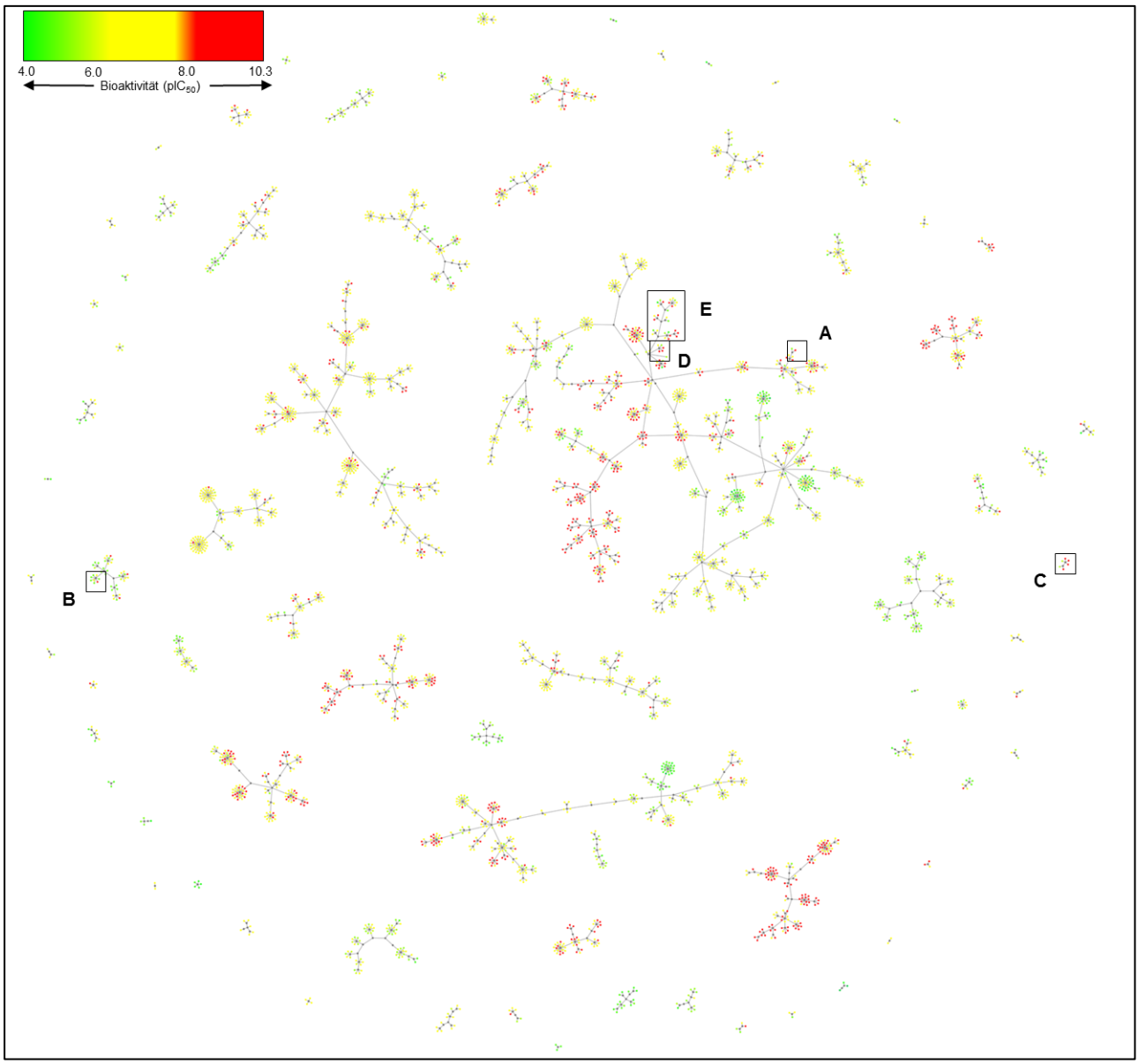


Abbildung 20.15. Das komplette inSARa-Netzwerkes des CDK2-Datensatzes (pIC_{50}) aus der BindingDB (Parameter: Mindest-MCS-Größe = 5 RG-Pseudoatome, Ausschlussliste = aktiv, Abbruchkriterium: ≤ 2% nicht-repräsentierte Moleküle). Automatisches Layout ohne Nachbearbeitung.

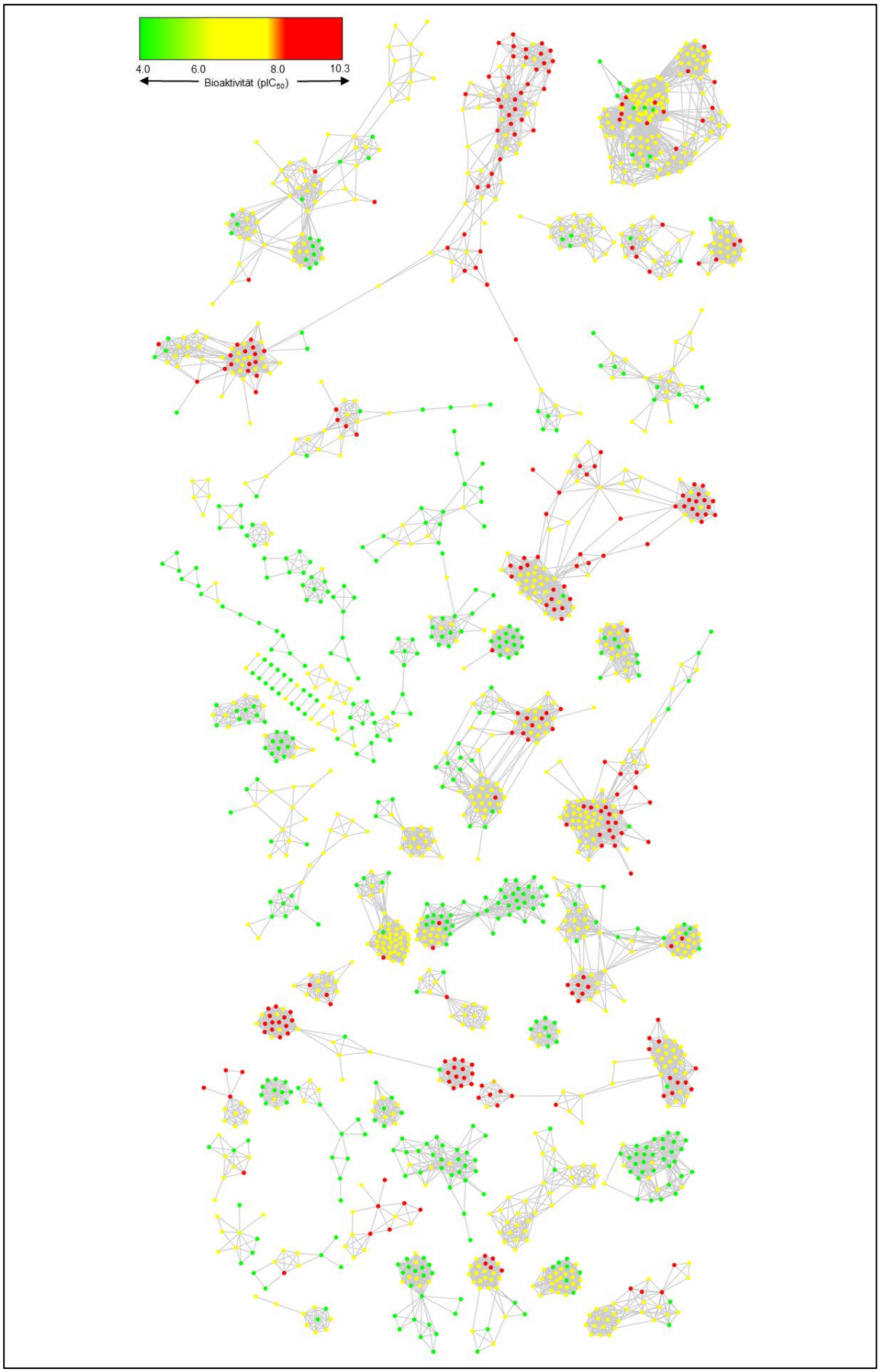


Abbildung 20.16. Fingerprint-basiertes Ähnlichkeits-Netzwerk für den CDK2-Datensatz (Fingerprint = ECFP4, Tc ≥ 0,55 für das Erstellen von Kanten). Das Netzwerk besteht aus 74 Komponenten.

Die nachfolgenden Fallbeispiele stammen aus den Subnetzwerk-Bereichen A-E des inSARa-Netzwerkes der CDK2 (Abbildung 20.16). Für einen allgemeinen Überblick über die ATP-Bindestelle bei Kinasen und typische Interaktionen sei auf ZUCCOTTO et al.[441] und ZHANG et al.[442] verwiesen.

Netzwerk-Ausschnitt A: Sprunghafte SARs

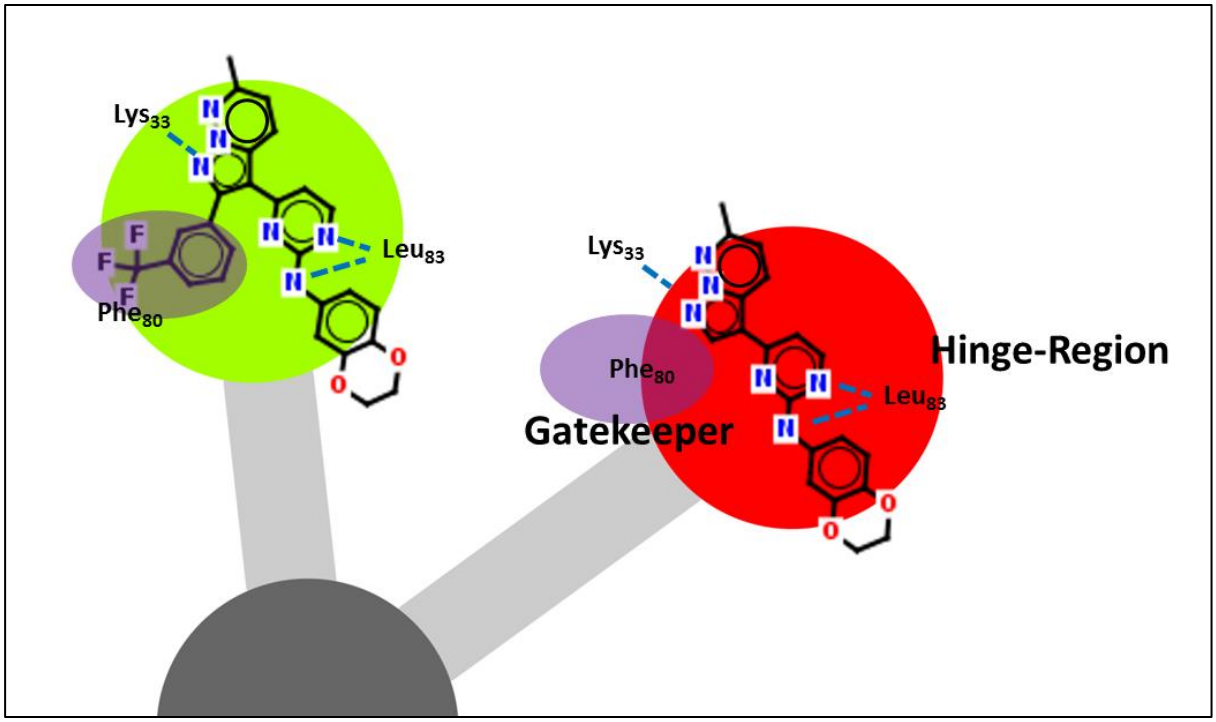


Abbildung 20.17. Sprunghafte SARs im Netzwerk-Ausschnitt A im CDK2-inSARa-Netzwerk. Trotz großem gemeinsamen MCS (Unterscheidung nur in lila markierter Trifluormethylphenylgruppe) weisen beide Moleküle eine große Bioaktivitätsdifferenz (pIC_{50} von 5,3 und 8,2) auf. Für Details siehe Text. Typische Aminosäuren der ATP-Bindetasche der CDK2 und potentielle H-Brücken-Interaktionen sind basierend auf Kristallstruktur-Informationen aus TAVARES et al.[443] und WYATT et al.[444] ergänzt. (Anmerkung: Bei einigen heterozyklischen bzw. annelierten Ringen kommt es aufgrund eines Bugs zu Fehlern bei der Molekülabbildung durch das verwendete Plugin chemViz. Der Ring, der die Aromatizität andeutet, wurde im annellierten Pyridazin-Ring daher in diese Abbildung manuell eingefügt.)

Obwohl sich beide N-Phenyl-4-pyrazolo[1,5-b]pyridazin-3-ylpyrimidin-2-amine[443] am MCS-Knoten A (Abbildung 20.17) nur durch den Trifluormethylphenyl-Rest unterscheiden, weisen sie einen Unterschied von fast drei Potenzbereichen in der Bioaktivität auf (sprunghafte SARs). Da es sich bei dieser Gruppe um einen sterisch anspruchsvolleren Rest handelt, könnte dieser drastische Aktivitätsverlust durch sterische Kollisionen verursacht sein. Die sterischen Voraussetzungen der hinteren Tasche (Gatekeeper-Region) werden bei der CDK2 durch das sperrige $Phenylalanin_{80}$ bestimmt.[441, 443] Nach Kristallstruktur-Daten für ein anderes Analogon des hochaktiven Moleküls liegt genau im Bereich des angefügten Restes diese Gatekeeper-Aminosäure, sodass die abgeleitete sterische Kollision eine sehr wahrscheinliche Erklärung ist.[444] Diese Erklärung wird auch durch ein anderes Molekülbeispiel mit vergleichbarem Verhalten von ZUCCOTTO et al.[441] untermauert.

Da die Trifluormethylgruppe aus 10 Nicht-Wasserstoff-Atomen besteht, ist der Größen-Unterschied der ausgetauschten Fragmente dieser Moleküle somit größer als 8 Nicht-Wasserstoff-Atome (AC-Kriterium mittels MMPA, vgl. Abschnitt 2.5.1), sodass dieses für die SAR-Interpretation wertvolle Molekülpaar mittels MMP-Analyse nicht als sprunghafte SARs erkannt würde. Dies unterstreicht, wie wertvoll die Flexibilität der MCS-basierte Analyse ist (vgl. Abschnitt 4.5).

Netzwerk-Ausschnitt B: SAR Hotspot

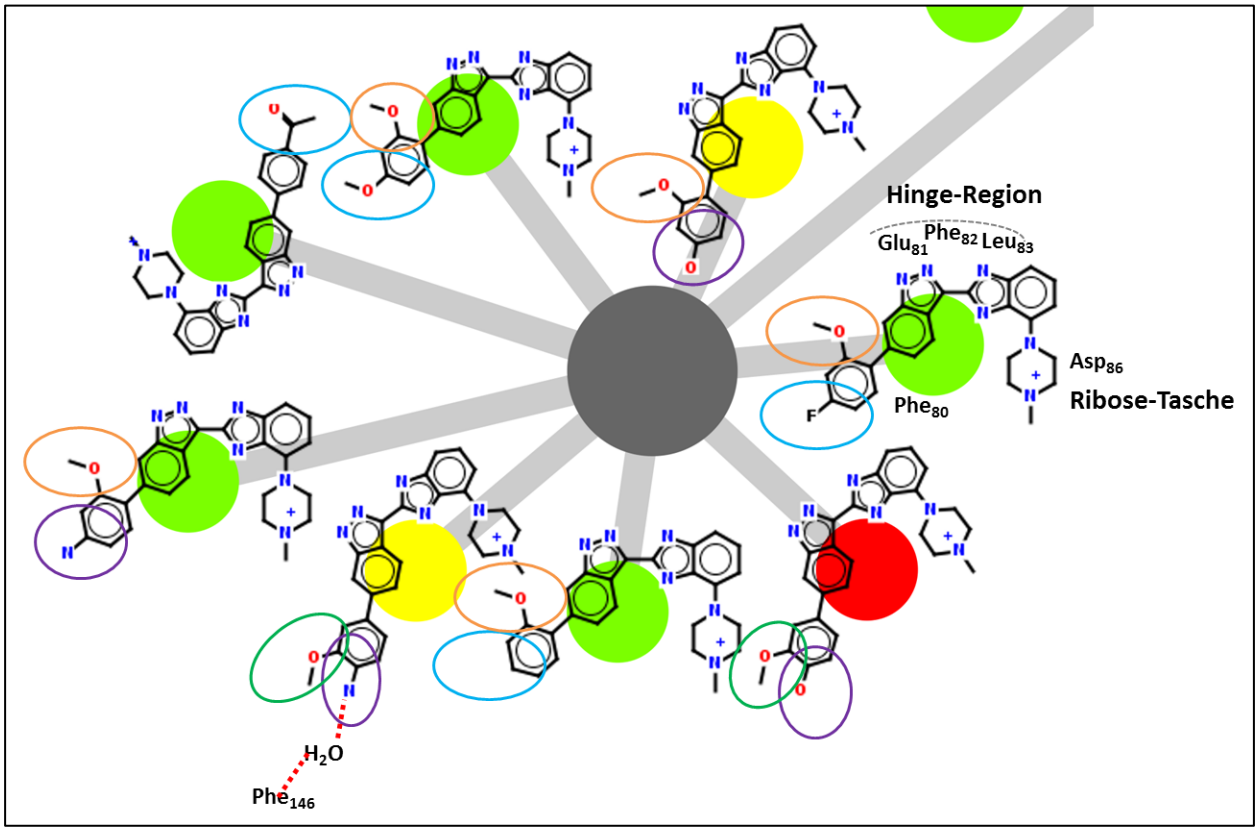


Abbildung 20.18. SAR Hotspot in Netzwerk-Ausschnitt B im CDK2-inSARa-Netzwerk. Die Bindetaschen-Informationen basieren auf Kristall-Strukturen aus der PDB (1EZV und 1EZR; CDK2). Das Substitutionsmuster, insbesondere der Substituent in 4- und 2-Position, am terminalen Phenylring beeinflusst entscheidend für die Bioaktivität (Bioaktivität variiert von µM bis nanomolar). Weitere Details siehe Text.

An dem MCS-Knoten in dem Netzwerk-Ausschnitt B (Abbildung 20.18) sind verschiedene 2-(6-Phenyl-1H-indazol-3-yl)-1H-benzo[d]imidazol-Analoga[445] zu finden, die deutliche Variation in der Bioaktivität aufweisen (SAR Hotspot). Es ist zu erkennen, dass das Substitutionsmuster am terminalen Phenylring entscheidenden Einfluss auf die Bioaktivität hat. Insbesondere die 4-Position und 2-Position sind von großer Bedeutung. Eine Methoxygruppe an 2-Position (braune Markierung) führt in allen Fällen zu einer Reduktion der Bioaktivität. Nur Moleküle mit einer HBD-Gruppe an 4-Position des Phenylringes (Hydroxyl- oder Aminogruppe, lila Markierung) weisen hohe oder mittlere Bioaktivität auf. Andere Substituenten (z.B. HBAs wie Carbonyl- oder Methoxygruppe) oder fehlende Substitution an 4-Position führen zu schwach aktiven Inhibitoren.

Eine mögliche Erklärung für die Bedeutung des HBD an 4-Position liefern die verfügbaren Kristallstrukturen für die 4-Amino-Analoga (PDB Code: 1EZV und 1EZR; CDK2). Eine H_2O-vermittelte Interaktion mit dem Phe-146 (1EZV) tritt bei dem 2-Methoxy-Analogon (1EZR) nicht auf (kein Wasser vorhanden, räumliche Verdrehung des Phenylringes, restliches Molekül unverändert) und stellt eine mögliche Erklärung für die höhere Bioaktivität der 3-Methoxy-Analoga dar.

Netzwerk-Ausschnitt C: Fehlgruppierung durch kleine RG-MCSs und unspezifische Strukturmotive

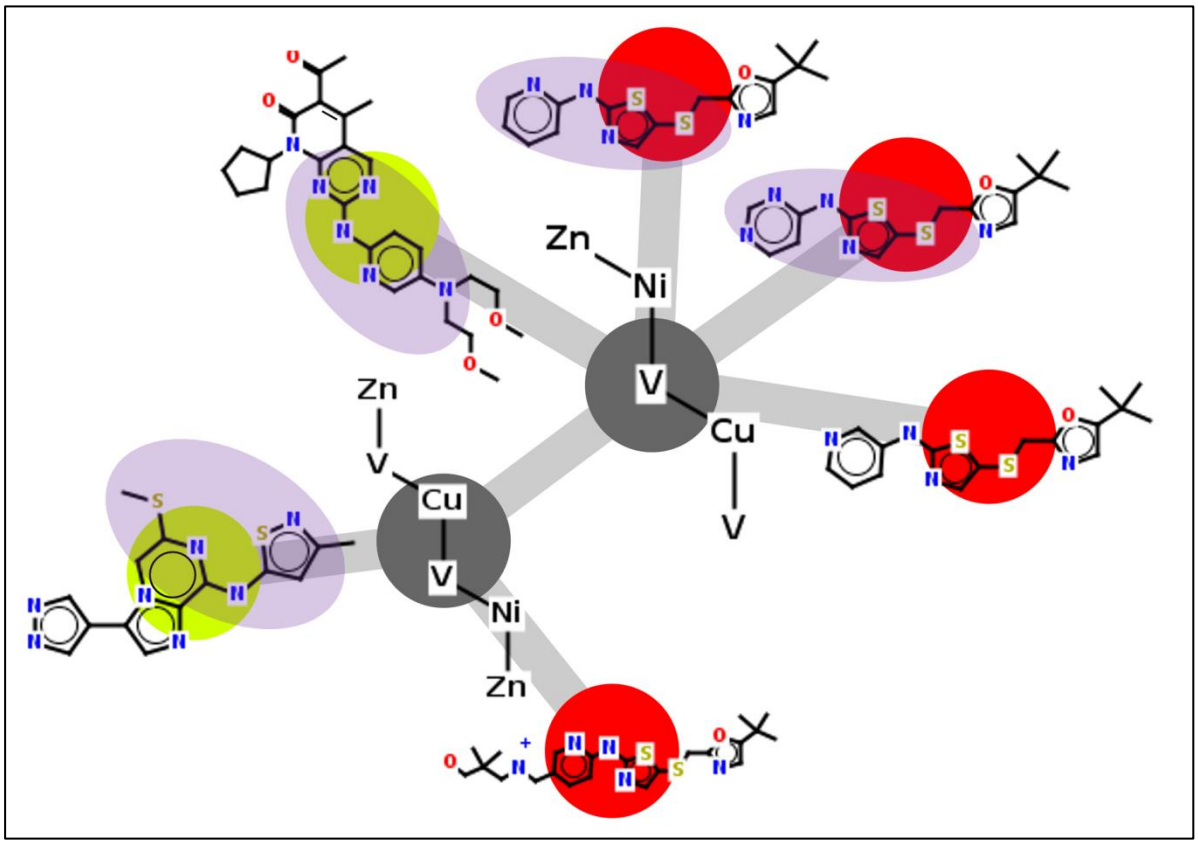


Abbildung 20.19. Beispiel für Fehlgruppierungen bzw. schwer interpretierbare Beziehungen an MCS-Knoten, die kleine RG-MCSs repräsentieren. Die durch die MCSs codierten Molekülteile sind lila markiert. Für Details siehe Text.

Die Kombination von roten und grünen Knoten in Netzwerk-Ausschnitt C ist ein wichtiger Hinweis auf potentielles Vorkommen von sprunghaften SARs. Bei genauerer Analyse (Abbildung 20.19) stellt man jedoch fest, dass die betroffenen MCS-Knoten relativ kleine MCSs (der Größe von 5 und 6 RG-Atomen) repräsentieren. Die zugehörigen Moleküle sind strukturell deutlich unterschiedlich, sodass es sich hierbei weder um sprunghafte SARs handelt noch wertvolle SAR-Information gefunden wird. Dies Beispiel zeigt, dass bei kleinen MCS-Knoten ein hohes Risiko für Fehlgruppierungen von Molekülen besteht.

Grund für diese schwer interpretierbare Fehlgruppierung ist das in Kinaseinhibitoren häufig vorkommende „Bisarylanilin“- bzw. „Heteroaryl-NH-Aryl“-Motiv[307, 446], das von beiden MCSs codiert wird („V-Cu-V“). Dieses Motiv ist wenig spezifisch, aber prädiktiv für Kinase-

inhibitorische Aktivität.[307] Grund dafür ist, dass dieses Strukturmotiv oder Elemente davon zumeist an der Bindung in der Hinge-Region (Nachahmung der H-Brückenbindungen, die normalerweise vom Adenin-Ring des ATPs mit der Kinase ausgebildet werden, vgl. PDB-Code 1QMZ; CDK2) beteiligt ist („Hinge-Bindungs-Motiv“: HBD, HBA und Aromat).[442]

Netzwerk-Ausschnitt D: Nutzen kleiner RG-MCSs zum Erkennen von minimal notwendigen Struktur-Elementen

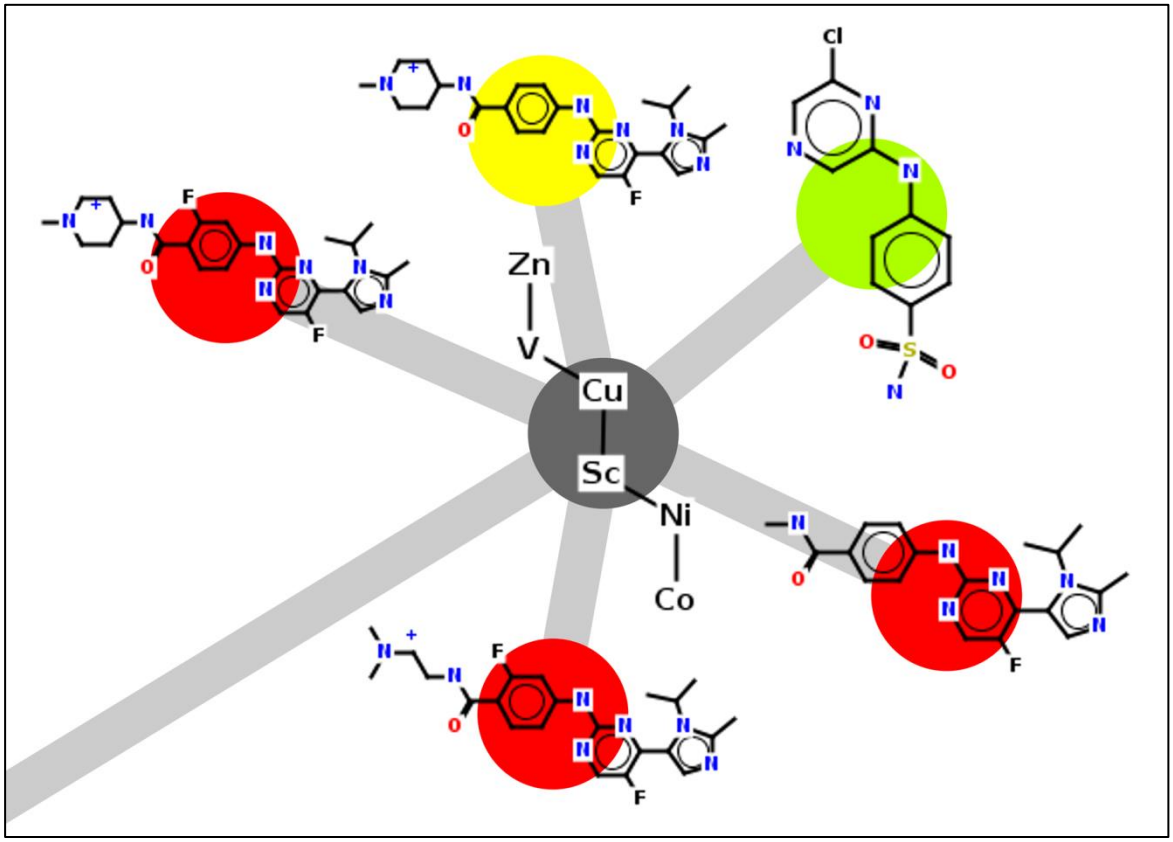


Abbildung 20.20. Beispiel für einen MCS-Knoten, der trotz der geringen Größes (6 RG-Atome) des repräsentierten MCS interessante Information für die SAR-Analyse codiert. Details siehe Text.

Der MCS-Knoten in Netzwerk-Ausschnitt D (Abbildung 20.20) zeigt, dass auch MCS-Knoten, die MCSs einer geringen Größe repräsentieren interessante Informationen liefern können. Der MCS entspricht dem RG des schwach aktiven Moleküls (grüner Knoten), während er bei den anderen, stärker aktiven Molekülen nur eine kleine Substruktur der jeweiligen RGs darstellt. Der MCS codiert in diesem Fall die strukturellen Voraussetzungen, die für eine Mindest-Bioaktivität am Target notwendig sind. Grund hierfür ist das typische „Hinge-Bindungs-Motiv“[442] (siehe oben). Bestätigt wird dies auch durch die verfügbare Kristallstruktur für dieses Fragment-Moleküls (vgl. PDB-Code: 2VTJ; CDK2), das aus einem Hochdurchsatz-Röntgenkristallographie-Screening einer Fragment-Bibliothek stammt[444]. Fragmente (Moleküle mit geringer Molekularmasse) weisen im Allgemeinen sehr geringe Bindungsaffinitäten (>100 µM) auf, wodurch hochsensitive biophysikalische Screening-Methoden wie die Röntgenkristallographie zur Messung verwendet werden.[444]

Durch Erweiterung dieses Minimal-Fragmentes am Cl-Ende in Richtung Gatekeeper-Region bzw. am (Sulfon)Amid-Ende in Richtung Ribose-Tasche kann die Bioaktivität durch die zusätzlichen Interaktionen in diesen Regionen deutlich gesteigert werden (vgl. PDB-Code 2W17; CDK2). Der hydrophobe Cl-Rest kann durch einen hydrophoben Hetero-Aromaten mit hydrophoben Alkyl-Gruppen ersetzt werden, sodass dieser hydrophobe Taschenbereich (Phe-80, Val-18) besser ausgefüllt wird (zusätzlich Interaktion mit Lys-33, vgl. 2W17). Die Sulfonamid-Gruppe kann durch ein Carbonsäureamid ausgetauscht werden und Linker plus eine PI-Gruppe (Vorteil: Steigerung der Löslichkeit[447]) kann angefügt werden, ohne dass die Interaktion mit dem Asp-86 verloren geht (vgl. 2W17 und 2VTJ).

Netzwerk-Ausschnitt E: Erkunden der chemischen Umgebung

Der Netzwerk-Ausschnitt E (Abbildung 20.21) soll die Stärke von inSARa Netzwerken im Vergleich zu den in Abbildung 20.16 (und auch Abbildung 20.12) dargestellten Fingerprint-basierten Netzwerken hervorheben.

Man sieht, dass durch die hierarchische Netzwerk-Struktur die interaktive SAR-Interpretation intuitiv ist und chemische Nachbarschaft sehr einfach durch Navigation durch das Netzwerk erkundet werden kann. Es ist deutlich zu erkennen, dass so leicht verwandte Scaffolds (vgl. gelber Kasten A in Abbildung 20.16) und Substitutionsmuster identifiziert werden können. Dies kann potentiell hilfreich für die Identifizierung von chemisch unerforschtem Raum sein.

Auch sieht man, dass zahlreiche Ideen für bioisosteren Austausch oder bioaktivitätserhaltende Molekülmodifikationen bzw. Anregungen für Scaffold-Hopping durch Netzwerk-Navigation generiert werden können. Des Weiteren werden leicht bioaktivitätsentscheidende (pharmakophore) Eigenschaften bzw. Strukturmerkmale deutlich.

Das Darstellen von Molekülen in verschiedener chemischer Umgebung ermöglicht gezielt unterschiedliche SAR-Aspekte wahrzunehmen. Im Fingerprint-basierten Netzwerk sind diese SAR-Beziehungen häufig auch zu finden, jedoch aufgrund der fehlenden klaren Netzwerk-Struktur und der unübersichtlichen Zahl an Verknüpfungen fällt es schwerer die entscheidenden Merkmale zu erfassen. Bei inSARa wird zumeist eine überschaubare Anzahl an Beziehungen an einem Knoten gezeigt, sodass die Interpretation weniger anspruchsvoll ist. Zudem helfen die zugehörigen MCSs bei dem Erkennen der gemeinsamen Merkmale, was die Interpretation ebenfalls unterstützt.

Durch das mehrfache Auftreten von Molekülen an kleineren oder größeren MCS-Knoten können in manchen Fällen auch die Schwächen der RG-Codierung ausgeglichen werden. Ein Beispiel hierfür stellt Molekül 1 in Abbildung 20.21 dar. Dadurch dass das Molekül 1 nicht nur am größten MCS-Knoten dargestellt wird, ist es möglich auch die Beziehung zu Molekül 2 (Cyclohexyl-Ring direkt statt über einen C1-Linker (=Hf-Zn-Ni) mit Ether-O verknüpft (=Hf-Ni) herzustellen. Durch den (fehlenden) Einschub des Zn-RG-Atoms trotz der gemeinsamen hydrophoben Eigenschaft der codierten Molekülreste wird der MCS aufgrund der notwendigen exakten Paarung verkleinert. Die generell schwer zu erfassenden hydrophoben Eigenschaften (vgl. Kapitel 7.4) sind insbesondere mit 2D-RGs schwer codierbar. Das Erkennen von gemeinsamen hydrophoben Eigenschaften führt auch bei inSARa (genauso wie auch bei 3D-Verfahren) häufiger zu Problemen.

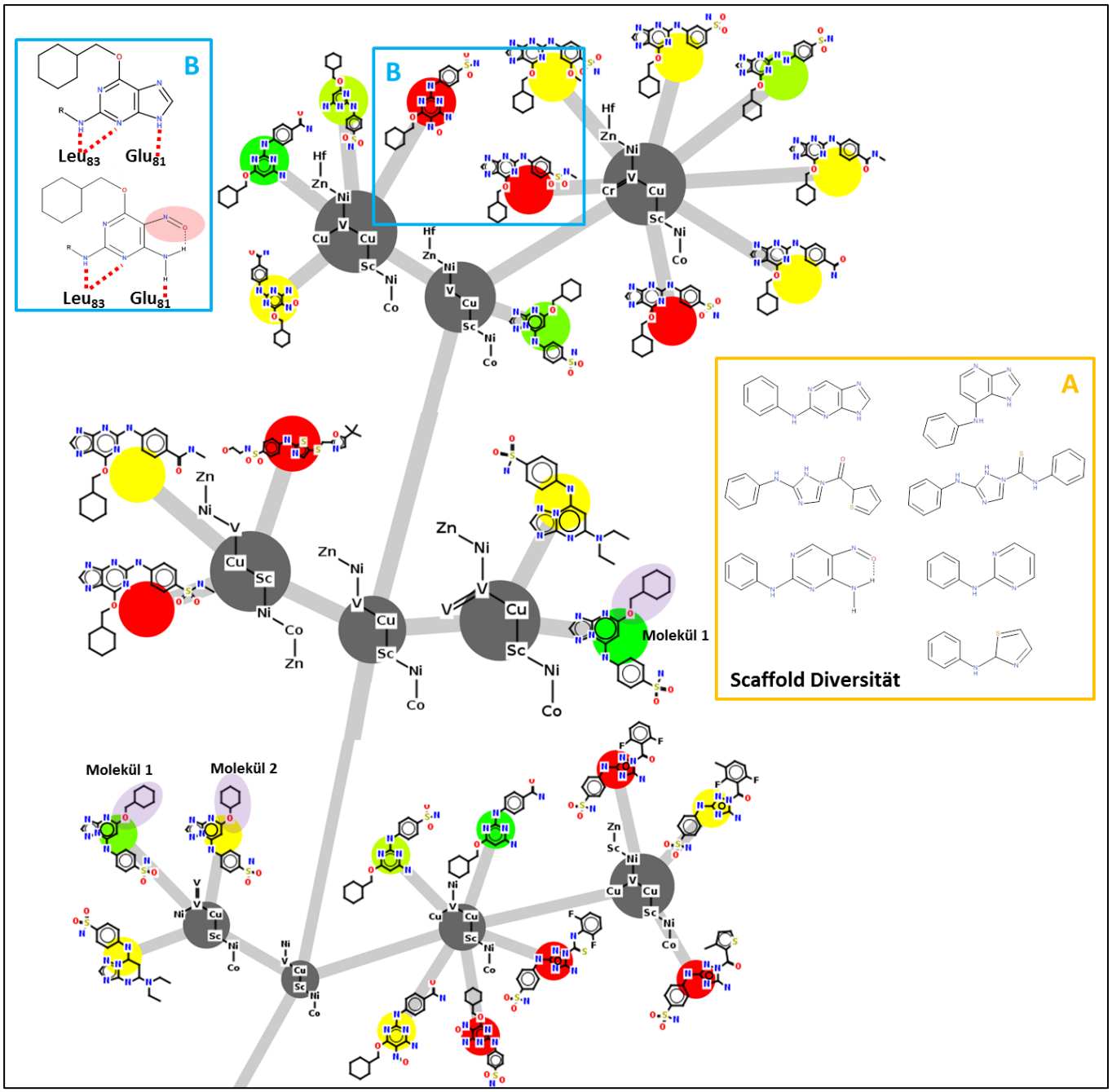


Abbildung 20.21. Erkundung der chemischen Umgebung durch Netzwerk-Navigation in inSARa-Netzwerken (am Beispiel des Netzwerk-Ausschnitt E im CDK2-Netzwerk). Verwandte Scaffolds (vgl. gelben Kasten A) und Substitutionsmuster werden schnell erkannt. Es können leichter (als im Fingerprint-Netzwerk) Beziehungen zwischen bestimmten Molekülen (hier: Nitroso-Pyrimidin- und Purin-Derivat, vgl. blauen Kasten B) erkannt werden. Dadurch dass das Molekül 1 nicht nur am größten MCS-Knoten dargestellt wird, ist es möglich auch die Beziehung zu Molekül 2 herzustellen. Weitere Details siehe Text.

Im blau markierten Bereich B in Abbildung 20.21 erkennt man zwei hochaktive, strukturell relativ ähnliche Moleküle (Nitroso-Pyrimidin- und Purin-Derivat) an zwei benachbarten Knoten. Der MCS-Knoten des Pyrimidin-Derivates zeigt, dass die Nitroso-Gruppe entscheidend für die hohe Bioaktivität ist. Durch Ausbilden einer intramolekularen H-Brücke im Nitroso-Derivat kann eine „Purin-mimetische“ Struktur bzw. Konformation (vgl. PDB-Code: 1OGU; CDK2) ausgebildet werden, sodass dieselben Interaktionen zur Hinge-Region wie im Purin-Derivat möglich sind (vgl. blauer Kasten B).[448] Man sieht, dass der Purin-Scaffold nicht

unbedingt für die Interaktion mit der CDK notwendig ist, sondern gegen strukturell vergleichbare Grundgerüste ausgetauscht werden kann.[448]

Es sollte jedoch bei der SAR-Analyse auch immer berücksichtigt werden, dass unter Umständen Assay-Artefakte in den Datensätzen vorhanden sind (vgl. Abschnitt 2.2.2). So gibt es beispielsweise auch bestimmte Nitroso-Derivate[449–450], die aufgrund ihrer Reaktivität kovalent mit Proteinen interagieren.

20.2.2. Beispiel: COX2

Nicht-steroidale Antiphlogistika gehören zu den am häufigsten eingesetzten Arzneistoffen.[451] Sie entfalten ihre antiinflamatorische, analgetische und antipyretische Wirkung zumeist durch unselektive Hemmung der Cyclooxygenase 1 und 2 (Abk. COX). Während die COX1 konstitutiv exprimiert wird und eine wichtige Funktion in vielen physiologischen Prozessen hat, wird die Isoform 2 (COX2) induktiv exprimiert und spielt eine wichtige Rolle v.a. bei pathologischen Prozessen wie Entzündung, Schmerz und Fieber und dysregulierter Proliferation.[452] Durch selektive COX2-Inhibitoren können die mit der gleichzeitigen Hemmung der COX1 einhergehenden unerwünschten gastralen und nephralen Wirkungen vermieden werden, wodurch sie wichtige Therapeutika inflammatorischer Erkrankungen darstellen.[451] Die COX2 gehört daher auch zu den stark beforschten Targets der letzten Jahrzehnte, wodurch nicht nur Kristallstruktur-Informationen und zahlreiche Bioaktivitätsdaten verfügbar sind, sondern auch wichtige pharmakophore Eigenschaften[453-455] aufgeklärt sind.

Eine große Gruppe an selektiven COX2-Inhibitoren weisen ein typisches strukturelles Motiv (sogenanntes „Mickey Maus“ Motiv, vgl. Abbildung 5.1) auf, das aus einem Diarylheterozyklus (4-, 5- oder 6-Ring ggf. mit Akzeptor-Eigenschaft) zusammen mit einer Sulfonylmethyl- bzw. Sulfonamid-Gruppe als HBA besteht (vgl. Abbildung 20.22).[452] Zu dieser Gruppe zählen auch das auf dem Markt befindliche Etoricoxib (Arcoxia®) und Celecoxib (Celebrex®) oder das aufgrund eines erhöhten kardiovaskulären Risikos zurückgerufene Rofecoxib (Vioxx®)[456].

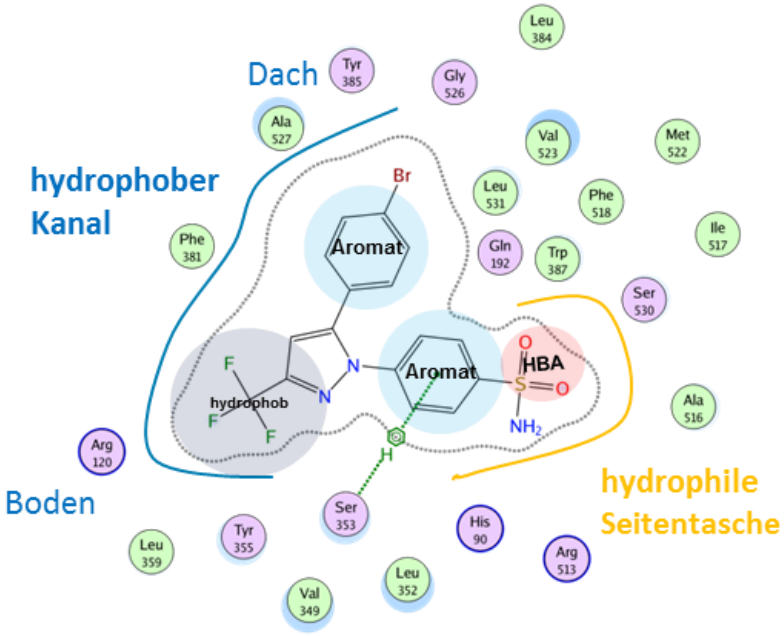


Abbildung 20.22. Typische COX2-Bindetasche-Merkmale und das klassische COX2-Pharmakophor am Beispiel eines Celecoxib-Derivates (Methylgruppe gegen Br-Atom ausgetauscht) nach MICHAUX et al.[454] (PDB-Code: 1CX2; COX2; Ligand-Interaktions-Diagramm erstellt mit MOE[188])

Der COX2-Datensatz (IC_{50}) aus der BindingDB besteht nach der Vorbereitung aus 2349 COX2-Inhibitor-Molekülen. In Abbildung 20.15 ist das resultierende inSARa-Netzwerk gezeigt (Einstellungen: Mindest-MCS-Größe = 4 RG-Atome, Abbruch-Kriterium: 2% nicht-repräsentierte Moleküle, Ausschlussliste = aktiv). Es besteht aufgrund der erhöhten Mindest-MCS-Größe aus 29 zusammenhängenden Komponenten, 61 Wurzel-Knoten und 1117 weiteren MCS-Knoten. 45 Moleküle sind nicht im Netzwerk repräsentiert. Aufgrund der geringeren Zahl bzw. der Größe der größten Komponente resultieren im automatischen Layout von Cytoscape einige Kantenkreuzungen. In Abbildung 20.24 ist zum Vergleich ein Fingerprint-basiertes Ähnlichkeits-Netzwerk (ECFP4-FP, Schwellenwert von Tc ≥ 0,55) dargestellt.

Eine Vielzahl der Moleküle im COX2-Datensatz weist ebenfalls das oben beschriebene Strukturmotiv auf. Der große Anteil heterogener Regionen im COX2-inSARa-Netzwerk, sowie der aus der SARI-Analyse resultierende heterogene SAR-Typ (globaler SAR-Index von 0.6, globaler SARI-Diskontinutitäts-Score von 0.7) belegt, dass dieses Motiv oftmals nur eine bestimmte Mindest-Aktivität an der COX2 garantiert. Für weitere Optimierung der Bioaktivität sind vielmals noch weitere Eigenschaften von Bedeutung (vgl. Netzwerk-Ausschnitte A und B in Abbildung 20.15) bzw. geringe strukturelle Variationen können starke Bioaktivitätsdifferenzen verursachen (vgl. Netzwerk-Ausschnitt C in Abbildung 20.15) wie die nachfolgenden zwei Beispiele veranschaulichen sollen.

Abbildung 20.23. Das komplette inSARa-Netzwerkes des COX2-Datensatzes (pIC50) aus der BindingDB (Parameter: Mindest-MCS-Größe = 4 RG-Pseudoatome, Ausschlussliste = aktiv, Abbruchkriterium: ≤ 2% nicht-repräsentierte Moleküle). Layout z.T. manuell nachbearbeitet.

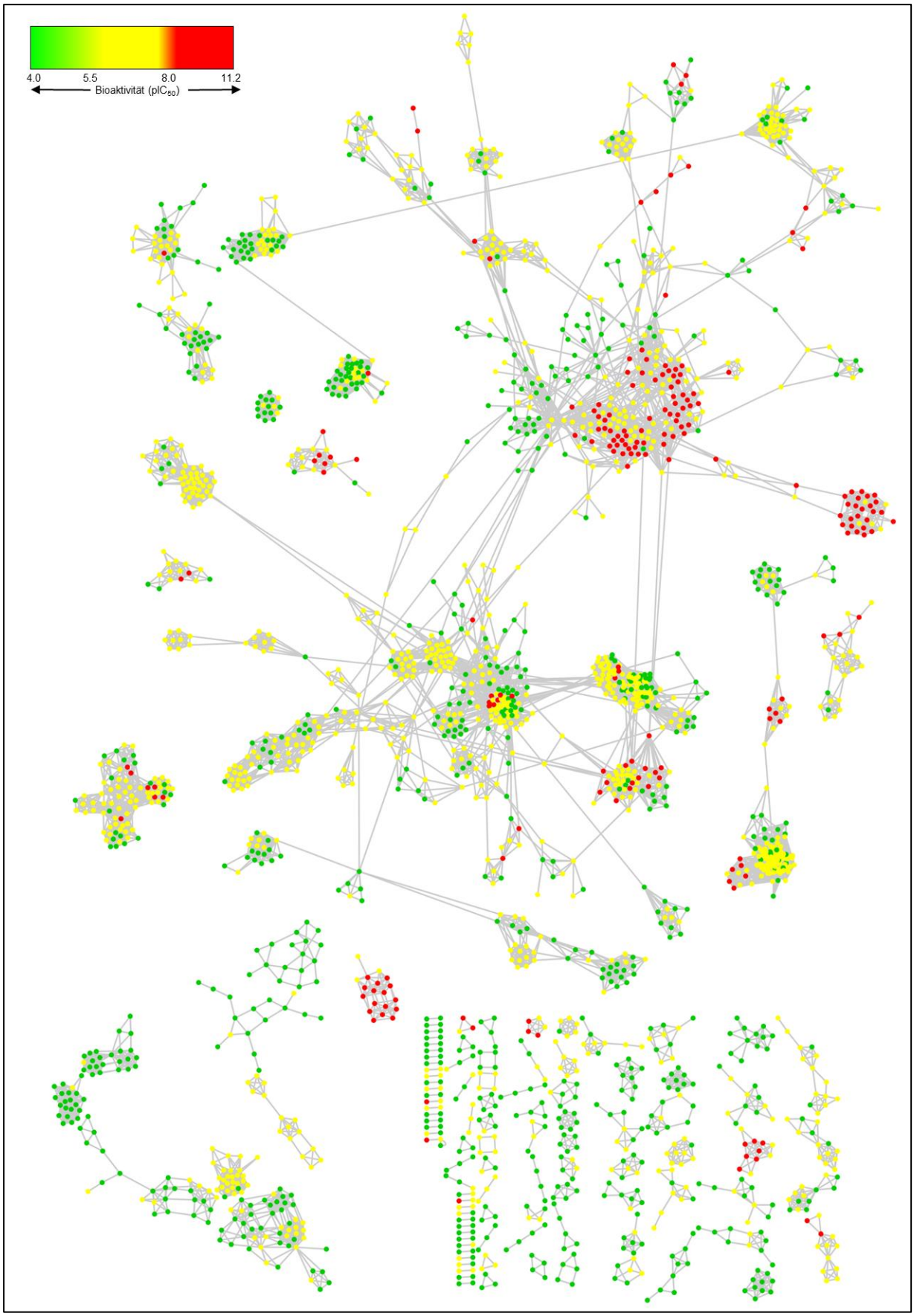


Abbildung 20.24. Fingerprint-basiertes Ähnlichkeits-Netzwerk für den COX2-Datensatz (Fingerprint = ECFP4, Tc ≥ 0,55 für das Erstellen von Kanten). Das Netzwerk besteht aus 115 Komponenten.

Netzwerk-Ausschnitt A und B: Weitere Typen von „Activity Switches“

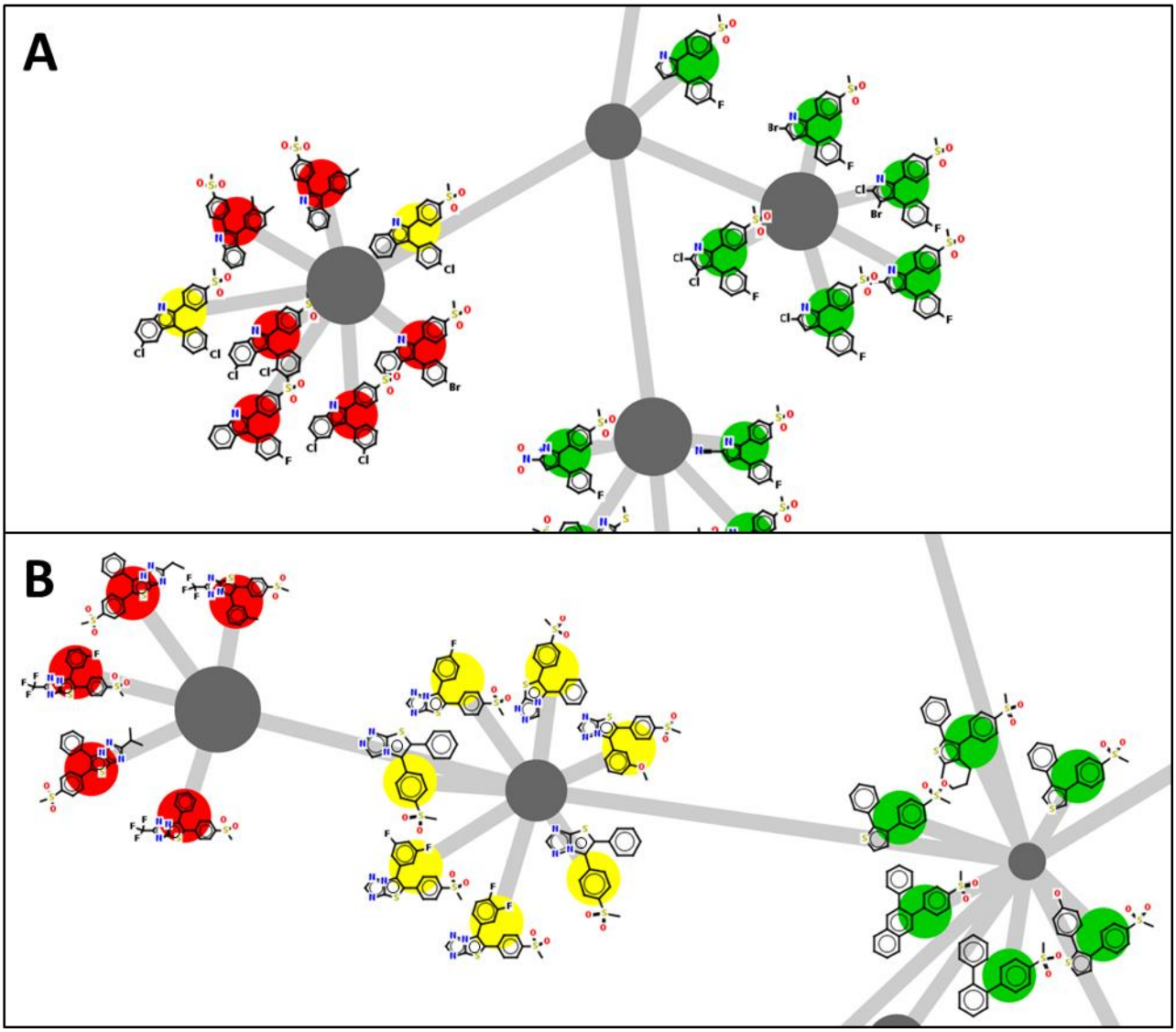


Abbildung 20.25. Verschiedene Typen von „Activity Switches“ im inSARa-Netzwerk der COX2. (A) Auftrennung von hoch (rot) und schwach aktiven (grün) Molekülen aufgrund unterschiedlicher pharmakophorer Eigenschaften. (B) „Ampel-Switch“: Pfad mit ansteigender Bioaktivität (grün-gelb-rot) durch Hinzufügen molekularer Eigenschaften. Weitere Details siehe Text.

In Abbildung 20.25 sind die Netzwerk-Ausschnitte A und B aus der Abbildung 20.15 im Detail dargestellt. Es sind verschiedene Typen von „Activity Switches“ erkennbar.

In Beispiel A ist zu erkennen, dass die Moleküle an den grünen Knoten nur schwach aktiv sind (IC_{50} > 1µM), obwohl das oben beschriebene strukturelle Motiv erfüllt wird. Eine Substitution des mittleren Pyrazol-Ringes sowohl mit hydrophoben Halogen-Atomen als auch mit verschiedenen HBAs führt zu keiner Aktivitäts-Steigerung. Erst die Annelierung eines aromatischen Phenylringes führt zu Molekülen mit hoher inhibitorischer Aktivität (pIC_{50} ≥ 8). Eine mögliche Erklärung für dieses Verhalten wären verbesserte hydrophobe Interaktionen am Boden des hydrophoben Kanals der Bindetasche (vgl. Abbildung 20.22) oder π-π Interaktionen des Phenylringes mit den aromatischen Aminosäuren Tyr-355 und Phe-381.

In Beispiel B ist der Ideal-Typ des „Activity Switches“ (sogenannter „Ampel-Switch“) zu sehen, der jedoch äußerst selten in Netzwerken zu finden ist. Die schwach aktiven Moleküle (grüne Knoten) weisen alle das klassische Grundmotiv (3 aromatische Ringe plus Sulfonylmethyl-Gruppe als HBA) auf. Durch Annelierung eines Heteroaromaten mit HBA/HBD-Funktion entstehen mittelaktive Derivate (gelbe Knoten). Weitere Addition von hydrophoben Gruppen (wie eine Trifluormethyl-, Ethyl- oder Isopropylgruppe) führt zu einer weiteren deutlichen Steigerung der Bioaktivität bis in den nanomolaren Bereich (rote Knoten). Auch sind verbesserte Ausfüllung der hydrophoben Tasche bzw. zusätzliche hydrophobe Interaktionen eine wahrscheinliche Erklärungen für dieses SAR-Verhalten.

Netzwerk-Ausschnitt C: Musterbeispiel für sprunghafte SARs

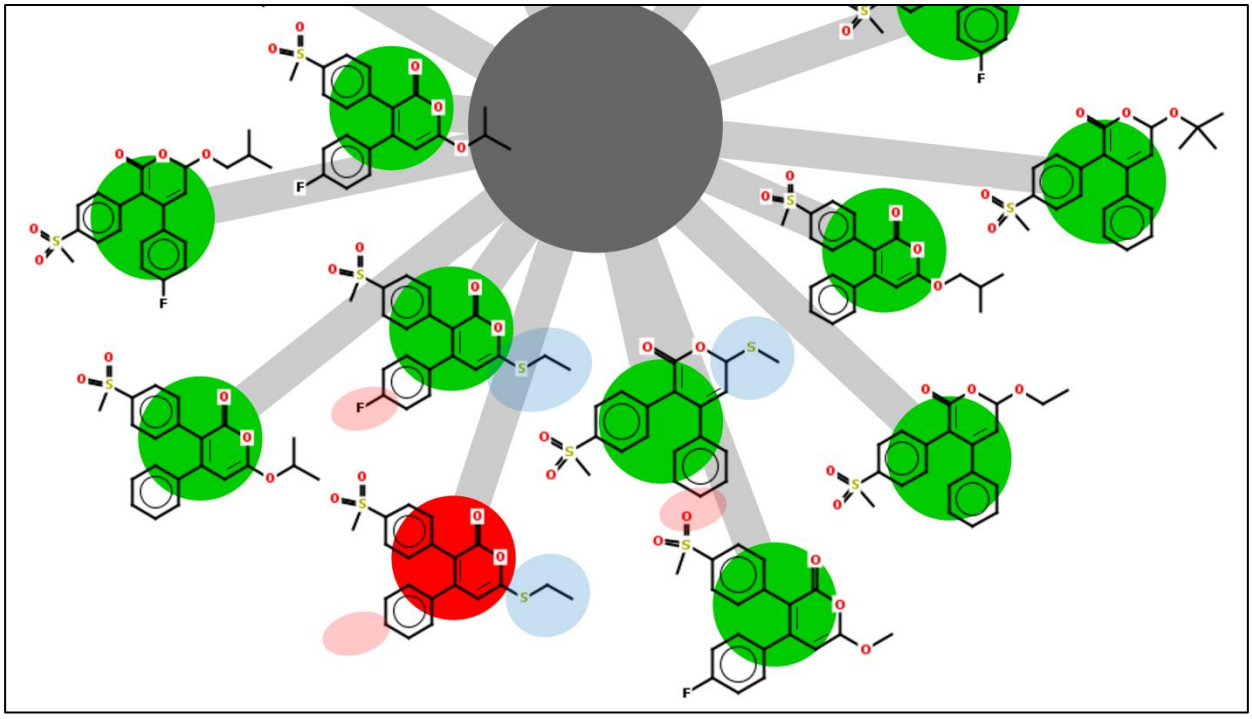


Abbildung 20.26. Musterbeispiel für sprunghafte SARs im inSARa-Netzwerk der COX2. Trotz geringer struktureller Unterschiede (vgl. rosa und blaue Markierungen) weist das Molekül, das vom roten Molekülknoten repräsentiert wird, deutlich höhere Affinität (IC_{50} = 3.2nM) zur COX2 auf als alle anderen Moleküle (IC_{50} > 1µM). Details siehe Text.

Alle Moleküle, die zu dem in Abbildung 20.26 gezeigten MCS-Knoten aus dem inSARa-Netzwerk des COX2-Datensatzes gehören, weisen eine große gemeinsame Substruktur auf. Nach dem SPP würde man somit erwarten, dass alle Moleküle eine ähnliche biologische Aktivität aufweisen. Mit Ausnahme des hochaktiven C6-Ethylthio-Derivates (vgl. roter Molekül-Knoten) trifft dieses Prinzip für alle weiteren Pyran-2-on-Derivate (geringe Bioaktivität, vgl. grüne Molekül-Knoten) zu. Dieses Molekül ist somit ein Musterbeispiel für unerwartete sprunghafte SARs. Erklären kann man sich dieses Verhalten ohne weitere Analysen von Protein-Ligand-Komplexen nicht, da alle Substituenten ähnliche elektronische und sterische Eigenschaften aufweisen. PRAVEEN et al. erklären dieses Verhalten mit der Ausbildung einer schwachen H-Brücke zwischen dem Schwefelatoms der Ethylthio-Gruppe und der OH-Gruppe des Serin-530, das sich in der hydrophoben Region befindet, in der die

Ethylthio-Gruppe beim Docking platziert wird.[457] Dies belegt abermals, wie schwer abschätzbar die Bedeutung von einzelnen H-Brückenbindungen ist (vgl. Kapitel 6.2) und wie wichtig optimales sterisches Einpassen in die Bindetasche (vgl. Kapitel 6.3) für die Bindungsaffinität ist. Wichtig bei der Analyse solcher Beispiele ist es, auch zu überprüfen, ob es sich nicht um einen Assay-Fehler oder Datenbank-Extraktions-Fehler (vgl. Kapitel 2.2.2) handelt. Manuelle Recherchen konnten in diesem Fall keine Inkonsistenzen dieser Art bestätigen: Der IC_{50}-Wert aus der BindingDB stimmt mit der Primärquelle überein. Laut Primärquelle ist der IC_{50}-Wert der Mittelwert aus einer Zweifachmessung mit weniger als 10% Abweichung vom Mittelwert.[457–458]

20.3. Zusammenfassung: SAR-Interpretation

Da inSARa-Netzwerke ohne Berücksichtigung von Bioaktivitätsinformation erzeugt werden, charakterisieren sie Datensätze basierend auf gemeinsam vorkommenden pharmakophorer Mustern. Obwohl die Bioaktivität keinen Einfluss auf die Gruppierung der Moleküle hat, konnte in den Beispiel-Netzwerken in vorangegangen Abschnitten gesehen werden, dass die Moleküle in vernünftiger Weise clustern, d.h. Moleküle mit ähnlicher Bioaktivität zusammengruppiert werden. Dies untermauert die Bedeutsamkeit der gewählten molekularen Repräsentation, also der RG-Definition. Denn bei der Wahl einer vernünftigen molekularen Repräsentation ist nach dem SPP ist zu erwarten, dass strukturell ähnliche Moleküle auch ähnliche biologische Aktivität besitzen. Auch unterstreicht die Tatsache, dass interpretierbare Beziehungen in den Netzwerken (wie anhand einer Vielzahl von Beispielen gezeigt) gefunden werden können, dass nicht nur die Art der molekularen Repräsentation, sondern auch des Ähnlichkeitsvergleiches sinnvoll ist.

inSARa-Netzwerke stellen eine neue Methode der SAR-Visualisierung und -Analyse dar. Im Gegensatz zu Fingerprint-basierten Ansätzen basieren diese Netzwerke auf klar-definierten Substruktur-Beziehungen unter Verwendung des intuitiven Konzepts des MCS. So ist jederzeit erkennbar, worauf die Ähnlichkeit zwischen gruppierten Molekülen beruht. Ein weiteres wichtiges Schlüsselmerkmal ist die hierarchische Netzwerk-Struktur, die (wie gezeigt) dazu beiträgt, dass sich die Interpretation geradlinig gestaltet. Größere oder kleinere gemeinsame Substrukturen sind in der Nachbarschaft zu finden, somit kann auch die chemische Nachbarschaft eines bestimmten Moleküls schnell erkundet werden. Durch das codieren pharmakophorer Eigenschaften (durch die Umwandlung der Moleküle in RGs) komplementiert inSARa andere Substruktur-basierte SAR-Analyse-Ansätze wie z.B. die BMMSGs, die auf MMP-Analyse basieren. Der Vorteil von inSARa im Vergleich zur MMPA ist eine größere Flexibilität im Erkennen von molekularen Gemeinsamkeiten. inSARa vereint die Vorteile der Substruktur- und RG-basierten SAR-Analyse und kann damit einige Probleme der jeweiligen Ansätze umgehen. So wird z.B. das Problem des exakten Substrukturabgleiches in MMP- oder MCS-basierten Ansätze sowie vordefinierter Einschränkungen (Größe und Anzahl auszutauschender Fragmente) in MMP-Ansätzen bei inSARa durch die RG-MCS-basierte Analyse umgangen.

Wie gezeigt, kann inSARa zur Analyse von Datensätzen verschiedener Größe (bis zu einigen tausend Molekülen) und struktureller Heterogenität verwendet werden. Der Grad der Netzwerk-Komplexität kann durch einige benutzerdefinierte Parameter gesteuert werden, wodurch ein Kompromiss zwischen Einfachheit bzw. Interpretierbarkeit und dem potentiellen

Verlust von SAR-Information möglich ist. Durch Definition einer Ausschlussliste, die aus der Analyse von Zufallsmolekülpaaren resultiert, kann unspezifische Ähnlichkeitsinformation von den Netzwerken ausgeschlossen werden.

Wie anhand der gezeigten Beispielsdatensätze zu sehen war, ermöglicht inSARa die Extraktion wichtiger SAR-Information von großen Datensätzen wie sie typischerweise in der Leitstruktur- oder „Hit-to-Lead“-Optimierungs-Stufe vorzufinden sind. Pharmakophore Muster, (nicht)bioisostere Austausche, „SAR Hotspots“ und „Activity Switches“ können identifiziert werden. Außerdem können sprunghafte SARs leicht erkannt werden, ohne dass ein vom verwendeten Fingerprint-abhängiger Schwellenwert definiert werden muss. In einer Vielzahl von Analysen werden sprunghafte SARs basierend auf paarweisen Molekülvergleichen analysiert (vgl. Abschnitt 2.5.1) ohne Berücksichtigung weiterer benachbarter Moleküle. Dadurch entsteht oftmals der Eindruck des häufigen Vorkommens von sprunghaften SARs, da ein einzelnes Molekül zum einen mit verschiedenen anderen strukturell ähnlichen Molekülen ebenfalls große Bioaktivitätsdifferenzen aufweisen kann oder aber Molekülpaare, die eigentlich Bestandteil eines SAR Hotspots sind, ebenfalls dieses Merkmal erfüllen können. Mit Hilfe von inSARa-Netzwerken können „echte“ sprunghafte SARs (Bioaktivitäts-Ausreißer in Umgebung von ansonsten kontinuierlichen SARs) sehr einfach von SAR Hotspots (Molekül befindet sich in heterogener Aktivitätslandschaft) unterschieden werden. Zudem wird ein AC basierend auf einer Gruppe von Molekülen erkannt, sodass die durch den Molekülpaar-basierten Vergleich resultierende Mehrfachzählung vermieden wird. Bei der Analyse einer Vielzahl verschiedener Datensätze ließ sich feststellen, dass in der Mehrheit SAR Hotspots in den Netzwerken zu finden sind. „Echte“ sprunghafte SARs stellen in den Netzwerken wie eigentlich zu erwarten ein eher seltenes Phänomen dar.

Durch das Zeigen von Molekülen in verschiedener chemischer Umgebung kann inSARa dazu beitragen Beziehungen (z.B. verschiedene Bindungsmodi wie bei FXa gezeigt oder Erklärung der hohen Affinität des Nitroso-Pyrimidinderivates durch Vergleich mit dem Purin-Analogon bei der CDK2) zu erkennen, die beispielsweise in Fingerprint-basierten Netzwerken nicht direkt ersichtlich sind. Wie bei der CDK2 veranschaulicht, ermöglicht die hierarchische Struktur des Netzwerkes, dass die interaktive SAR-Interpretation intuitiv ist und die chemische Nachbarschaft leicht durch Navigation durch das Netzwerk erkundet werden kann. So können verwandte Scaffolds und Substitutionsmuster leicht erkannt werden. Dies kann möglicherweise hilfreich sein bei der Identifizierung von bisher nicht-erkundetem oder wenig erkundetem, aber vielversprechendem chemischen Raum, woraus Ideen für neue Arzneistoffkandidaten-Projekte resultieren können.

21. Ergebnisse und Diskussion: inSARa Hybrid

21.1. Variante A: Cluster-Analyse

Anhand des Thrombin-Datensatzes soll im Folgenden exemplarisch das Prinzip der Variante A des InSARa Hybrid Ansatzes aus Kapitel 14.2 demonstriert werden.

Der Thrombin-Datensatz aus Tabelle 11.1 stellt mit fast 3000 Molekülen einen sehr großen Datensatz dar. Das resultierende inSARa-Netzwerk unter Verwendung aller Moleküle wird folglich sehr groß und komplex (vgl. Tabelle 18.1 und Tabelle 18.3 in Kapitel 18.3). Neben den in Kapitel 18.3 erläuterten Optionen zur Reduktion der Netzwerk-Komplexität stellt die Analyse einzelner Komponenten des Fingerprint-basierten Ähnlichkeits-Netzwerkes eine weitere Variante der SAR-Analyse dar.

In Abbildung 21.1 wird das Fingerprint-basierte Ähnlichkeits-Netzwerk für den kompletten THR-Datensatz unter Verwendung des ECFP4-Fingerprints und einem Schwellenwert $Tc \geq 0.55$ gezeigt. Das Netzwerk besteht aus einer sehr großen Komponente, die 968 Moleküle repräsentiert, ein paar mittelgroßen Komponenten, die etwa 100 bis 300 Moleküle repräsentieren, und einer Vielzahl kleiner Komponenten. Wie in Tabelle 21.1 dargestellt, kann über die Variation des Tc-Schwellenwertes die Topologie des Netzwerkes beeinflusst werden. Je höher der Schwellenwert gewählt wird, desto mehr Komponenten entstehen und desto kleiner werden die einzelnen Komponenten.

Tabelle 21.1. Einfluss des Tc-Schwellenwertes auf die Topologie des Fingerprint-basiertes Ähnlichkeits-Netzwerk für den THR-Datensatz (Fingerprint = ECFP4).

Tc-Schwellenwert	0.5	0.55	0.6
Gesamtzahl an Komponenten	65	84	119
Zahl der Moleküle in größter Komponente	1023	968	705
Zahl der Moleküle in zweitgrößter Komponente	316	284	284
Zahl der Moleküle in drittgrößter Komponente	184	155	152
Zahl der Moleküle in viertgrößter Komponente	172	121	107

Die größte Komponente des Netzwerkes (blau markiert) aus Abbildung 21.1 ist sehr groß und zum Teil sehr heterogen bezüglich der Bioaktivität. Aufgrund der fehlenden klaren Struktur des Netzwerkes ist die SAR-Interpretation folglich relativ schwierig. Eine Möglichkeit der Analyse dieses Cluster stellt, die Erstellung eines inSARa-Netzwerkes auf Grundlage der zugehörigen Moleküle dar. Das resultierende Netzwerk (für Mindest-MCS-Größe = 6 RG-Atome und Abbruchkriterium ≤ 5% nicht-repräsentierte Moleküle) ist in Abbildung 21.2 dargestellt. Es besteht aus insgesamt 867 MCS-Knoten (davon 9 Wurzel-Knoten), 2 Komponenten und repräsentiert insgesamt 927 Moleküle (95,9% der Moleküle der Komponente 1).

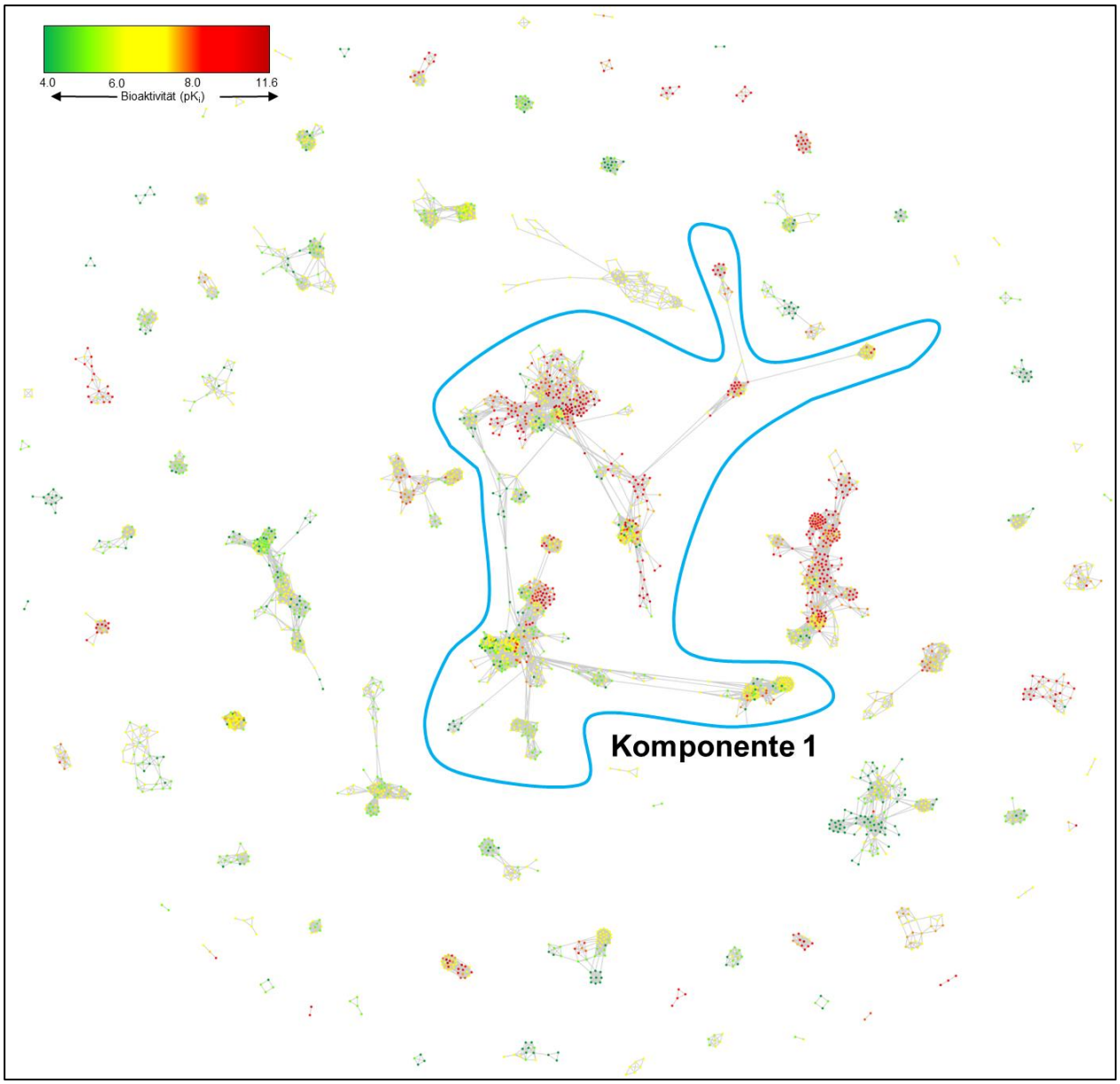


Abbildung 21.1. Fingerprint-basiertes Ähnlichkeits-Netzwerk für den THR-Datensatz (Fingerprint = ECFP4, Tc ≥ 0.55 für das Erstellen von Kanten). Das Netzwerk besteht aus 84 Komponenten.

Abbildung 21.2. inSARa-Netzwerk der größten Komponente (Komponente 1 in Abbildung 21.1) des THR-Datensatzes (pK$_i$) aus der BindingDB (Parameter: Mindest-MCS-Größe = 6 RG Pseudoatome, Ausschlussliste = aktiv, Abbruchkriterium: ≤ 5% nicht-repräsentierte Moleküle). Layout manuell nachbearbeitet.

Im Vergleich zum inSARa-Netzwerk basierend auf dem gesamten Datensatz ist das Netzwerk deutlich weniger komplex (vgl. Tabelle 18.3) und somit leichter überschaubar. Im Vergleich zum Fingerprint-basierten Netzwerk weist das inSARa-Netzwerk eine klare Struktur auf und die für die Gruppierung verantwortlichen pharmakophoren Eigenschaften sind sofort ersichtlich. Die SAR-Interpretation ist somit deutlich einfacher. Auch lässt sich feststellen, dass auf Basis gemeinsamer pharmakophorer Eigenschaften Moleküle ähnlicher Bioaktivität zumeist zusammengruppiert werden. So lassen sich in Abbildung 21.2 klar Bereiche abgrenzen, wo v.a. Moleküle hoher Bioaktivität (rote Umrandung), mittlerer Bioaktivität (gelbe Umrandung) oder schwacher bis mittlerer Bioaktivität (grün-gelbe Umrandung) gefunden werden. Es lassen sich jedoch auch Bereiche finden, wo die Bioaktivität der Moleküle an einem MCS-Knoten stark variiert (vgl. z.B. rot-gelb-grün umrandeten Bereich). Auch finden sich in den homogeneren Bereichen immer wieder Knoten, wo einzelne Moleküle sich in der Bioaktivität deutlich von allen übrigen Molekülen des jeweiligen MCS-Knoten unterscheiden. Dies ist im Einklang mit den Ergebnissen der inSARa^{auto}-Analyse aus Kapitel 22.1, wo für den THR-Datensatz ein großer Anteil homogener Bereiche, aber auch viele SAR Hotspots identifiziert wurden.

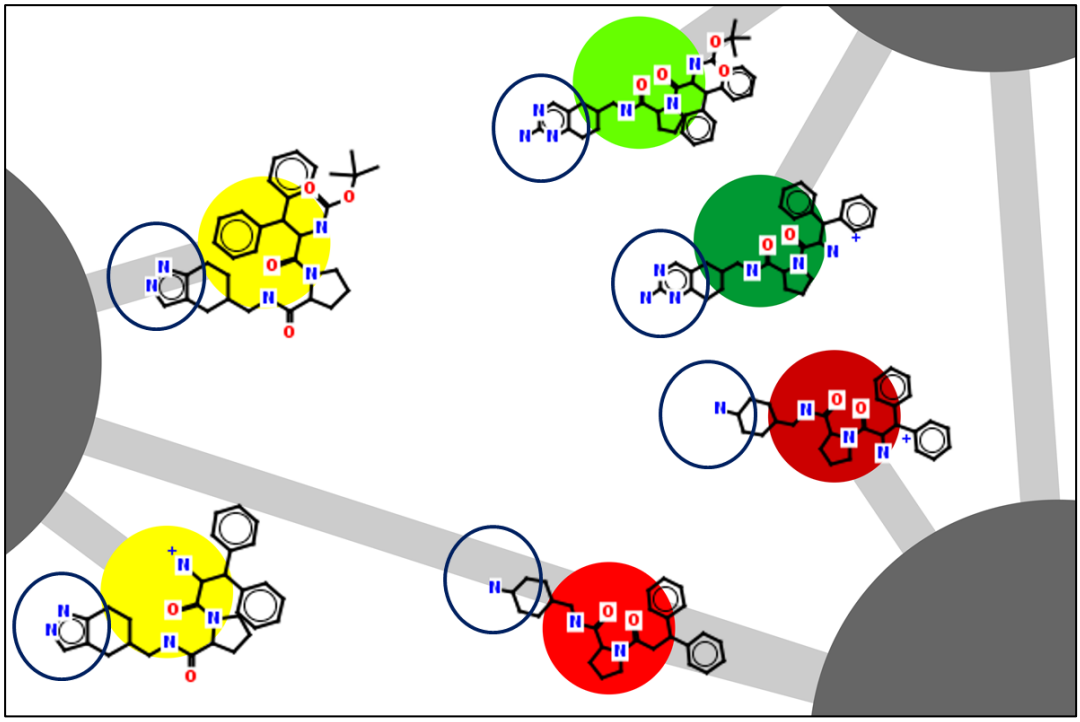


Abbildung 21.3. „Aktivitäts-Dreieck“ im inSARa-Netzwerk aus Abbildung 21.2. MCS-Knoten mit jeweils schwachaktiven (pK_i zwischen 4.0 und 6.0, grün), mittelaktiven (pK_i zwischen 6.0 und 8.0, gelb) und sehr hochaktiven Molekülen (pK_i > 10, (dunkel)rot) in direkter Nachbarschaft. Blaue Umrandung der aktivitätsentscheidenden Gruppen.

Anhand des inSARa-Netzwerkes lassen sich schnell aktivitätsentscheidende Merkmale wie z.B. anhand des in Abbildung 21.3 gezeigten auffälligen „Aktivitäts-Dreiecks“ identifizieren. Die blau umrandeten Gruppen am Cyclohexyl-Ring sind aktivitätsentscheidend. Analog zu FXa ist aufgrund der hohen strukturellen Ähnlichkeit bei der Serinprotease Thrombin ebenfalls eine ionische Wechselwirkung einer basischen P1-Gruppe eine Schlüssel-

Interaktion mit dem Asp_{189} in der S1-Tasche (sogenannte „Selektivitätstasche“)[459]. Das primäre aliphatische Amin weist von allen Gruppen die höchste Basizität auf (d.h. optimale elektronische Voraussetzung für die Interaktion) und scheint sterisch ebenfalls optimal in die „Selektivitätstasche“ zu passen. Das Pyrazol und das 2-Aminopyrimidin sind deutlich schwächere Basen, zudem sind die Gruppen sterisch anspruchsvoller, was scheinbar einen deutlichen Verlust an Bioaktivität verursacht.

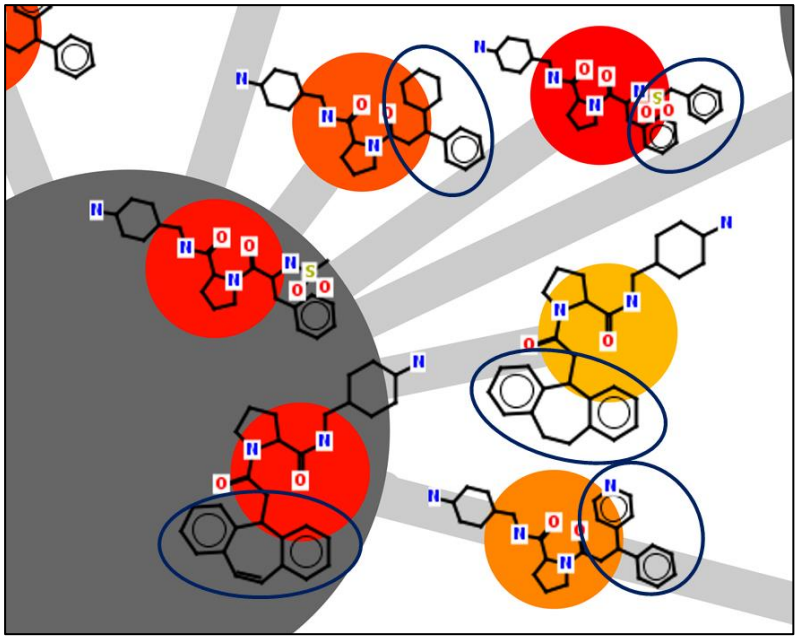


Abbildung 21.4. Beispiel für bioisosteren Austausch der lipophilen P3-Gruppe (blau markiert) unter Erhalt der hohen Bioaktivität.

Anhand des in Abbildung 21.4 dargestellten MCS-Knoten ist zusehen, dass im inSARa-Netzwerk sehr leicht Ideen für bioisosteren Austausch erhalten werden können. Die beiden in der lipophilen S3-Tasche (z.B. Trp_{215})[459–460] bindenden Phenylringe können sowohl verbrückt als auch gegen einen Pyridin oder nicht-aromatischen lipophilen Cyclohexylring bioisoster unter Erhalt der hohen Bioaktivität ausgetauscht werden (siehe blaue Markierung).

Im Folgenden sollen einige Beispiele für MCS-Knoten mit Molekülen, die deutliche Unterschiede in der Variabilität zeigen, untersucht werden, um zu analysieren, ob es sich um „künstliche“ (d.h. aufgrund von Fehlgruppierungen oder zu starker molekularer Abstraktion) oder „echte“ ACs oder SAR Hotspots handelt.

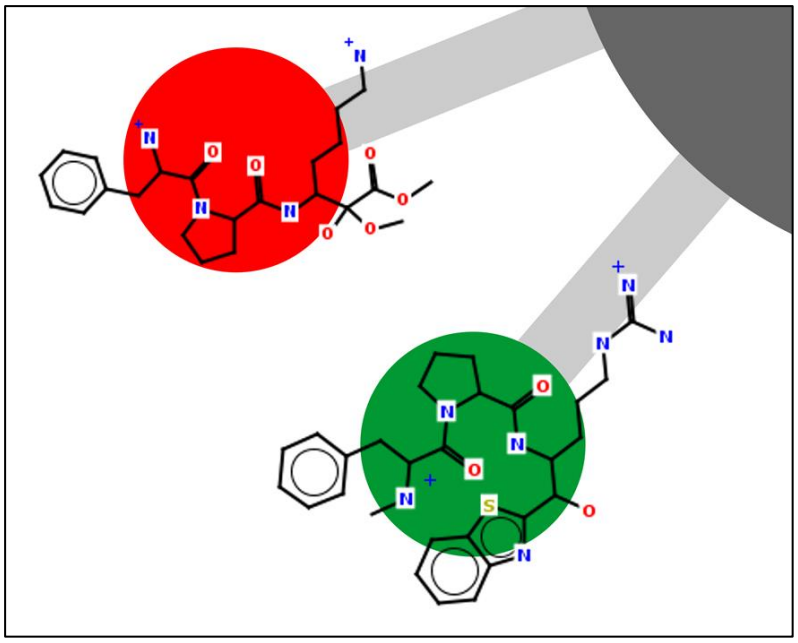


Abbildung 21.5. Potentiell sprunghafte SAR: Hochaktives (pK_i > 8, rot) und sehr schwachaktives Molekül (pK_i < 5, dunkelgrün) an einem großen terminalen MCS-Knoten (MCS-Größe 11 RG-Pseudoatome).

In Abbildung 21.5 ist ein potentieller AC-MCS-Knoten mit zugehörigen Molekülen dargestellt. Aufgrund der MCS-Größe würde man hier ein Beispiel für sprunghafte SARs erwarten. Die Moleküle weisen jedoch trotz großer gemeinsamer Substruktur noch einige strukturelle Unterschiede auf, sodass es sich hierbei nicht um ein AC handelt. Grund hierfür ist die im Vergleich zu anderen Datensätzen (vgl. Abbildung 26.1 im Anhang und Tabelle 18.2 in Kapitel 18.3) erhöhte Größe der RGs (bei den gezeigten Molekülen 14 RG-Pseudoatome). Die Ursache für die großen RGs ist, dass es sich bei den meisten Thrombin-Inhibitoren im Datensatz um Peptidomimetika handelt. Die ersten, selektiven Thrombin-Inhibitoren (z.B. PPACK) waren zumeist Tripeptide mit der Sequenz D-Phe-Pro-Arg[461]. Dies wird ebenfalls in dem schwachaktiven Molekül imitiert. Das hochaktive Molekül ist hingegen ein Phe-Pro-Lys-Mimetikum. Ein Grund für die schwache Bioaktivität könnte z.B. der sterisch anspruchsvollere Benzothiazol-Substituent sein. Das Hemiacetal in Nachbarschaft zur Esterfunktion in dem hochaktiven Molekül ist jedoch ebenfalls relativ voluminös. Das Besondere hierbei ist die Nachahmung des Übergangszustandes bei der Hydrolyse der Peptidbindung. Die Alkoholfunktion ist im Gegensatz zu einigen hochaktiven elektrophilen alpha-Keto-Benzothiazolen (Aktivierung des Kohlenstoff-Atoms durch die Carbonylgruppe) kein Übergangszustand-Analogon.[459] Da die MCS-Sim für diesen Knoten < 0.8 ist, würde dieser Knoten von inSARa^{auto} ebenfalls nicht als AC klassifiziert werden, sondern einen Knoten mit heterogener SAR-Information (weder AC noch SAR Hotspot) darstellen.

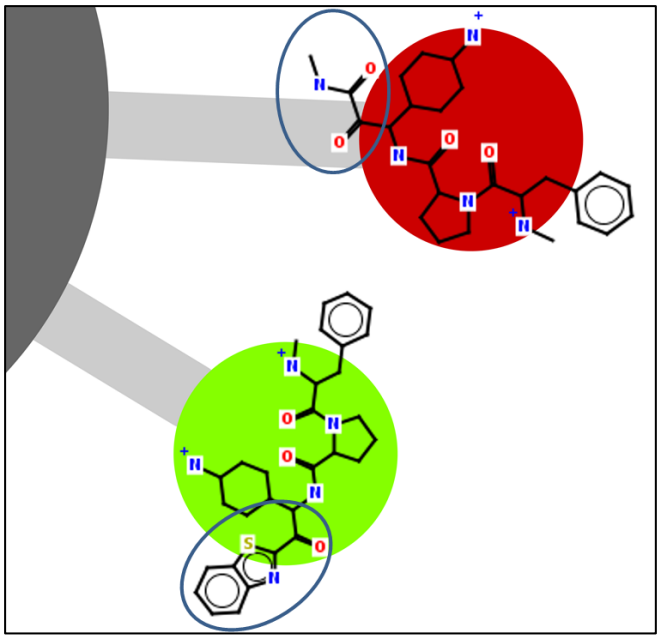


Abbildung 21.6. Weiteres Beispiel für eine potentielle sprunghafte SAR: Sehr hochaktives (pK_i > 10, dunkelrot) und schwachaktives Molekül (pK_i < 6, grün) an einem großen terminalen MCS-Knoten (MCS-Größe 12 RG-Pseudoatome). Struktureller Unterschied bzw. aktivitätsentscheidende Gruppe ist blau markiert.

In Abbildung 21.6 ist ein weiterer potentieller AC-MCS-Knoten gezeigt. Dieser weist einen MCS-Sim von 0.86 auf, sodass dieser Knoten von inSARa^{auto} als AC klassifiziert würde. Beide Moleküle sind strukturell sehr ähnlich mit Ausnahme des blau markierten Strukturelements. Das hochaktive Molekül weist eine hochreaktive β-Ketoamid-Struktur auf, das aufgrund der elektrophilen Ketofunktion typischerweise kovalente Bindungen mit nucleophilen Aminosäuren (z.B. Ser_{195} in Thrombin) eingeht.[462] Dies ist eine einfache Erklärung für diese sprunghafte SAR.

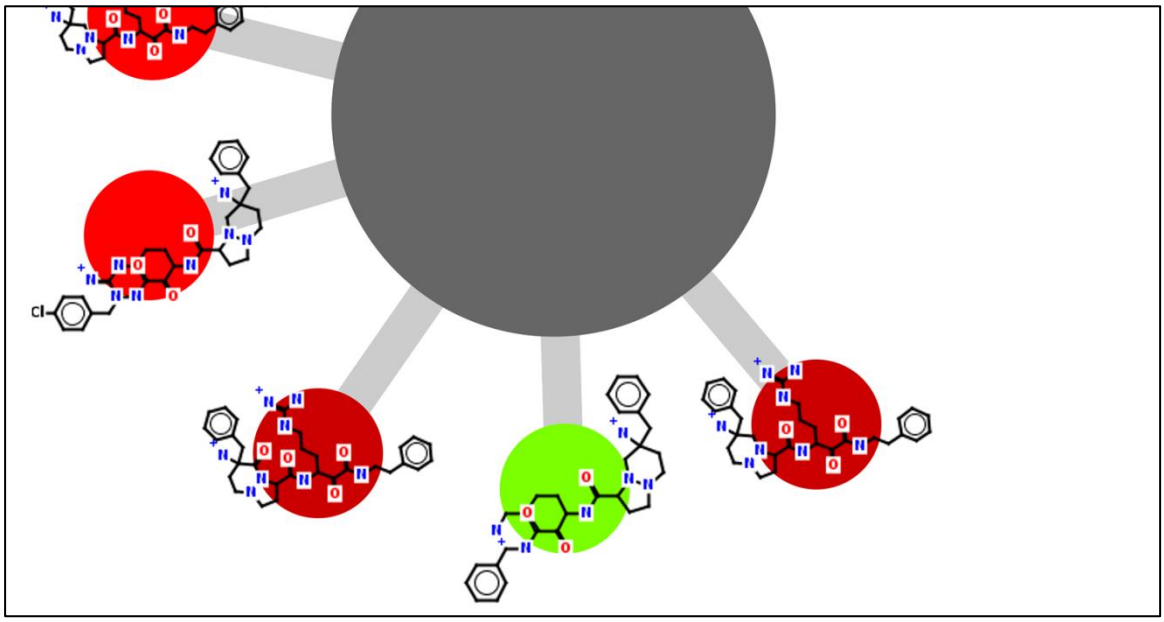


Abbildung 21.7. Sprunghafte SARs: Schwachaktives Molekül (grün) in Umgebung von hochaktiven Molekülen (dunkelrot). Details siehe Text.

Abbildung 21.7 zeigt ein weiteres Beispiel für sprunghafte SARs. Aufgrund der molekularen Abstraktion weisen das grüne und das rechte dunkelrote Molekül den gleichen RG auf. Die Moleküle weisen alle eine PI-Gruppe auf (Guanidin oder primäres aliphatisches Amin), die eine Schlüsselinteraktion mit der S1-Tasche eingeht (siehe oben). Bei den hochaktiven Molekülen ist die PI-Gruppe jedoch über einen C3-Linker, bei dem schwachaktiven über einen C4-Linker verknüpft (im RG jeweils durch Zn codiert). Zudem ist der Phenylring über einen C2-Linker in den Hochaktiven verknüpft, beim Schwachaktiven findet sich ein C1-Linker. Der Vorteil der RG-Abstraktion ist, dass aufgrund der Einheitscodierung von Linker-Abständen und pharmakophoren Eigenschaften (PI-Gruppe) diese Moleküle an einem MCS-Knoten zusammengruppiert werden und so die Wichtigkeit des Abstandes zwischen bestimmten pharmakophoren Gruppen erkannt werden kann.

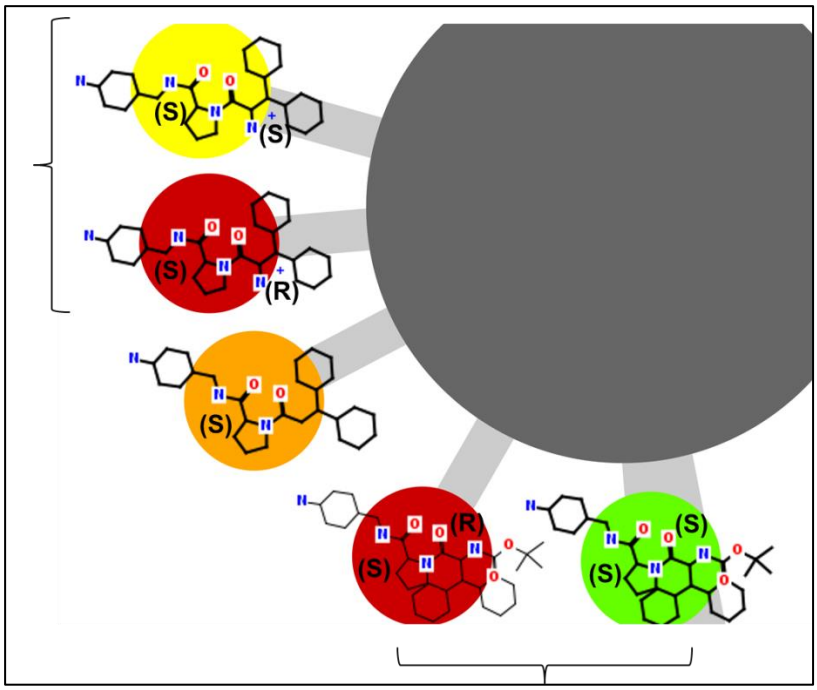


Abbildung 21.8. SAR Hotspot im Thrombin-inSARa-Netzwerk: Bedeutung der Stereochemie für die Bioaktivität. Details siehe Text.

Abbildung 21.8 zeigt einen SAR Hotspot aus dem inSARa-Netzwerk. Anhand der markierten Molekülpaare (Klammern) ist schnell das bioaktivitätsenscheidende Merkmal, also die Stereochemie an dem chiralen Kohlenstoff-Atom der P3-Gruppe[460], identifiziert. (S)-Konfiguration führt zu einem deutlichen Bioaktivitätsverlust. Obwohl Stereochemie in RGs nicht codiert ist, ist es mit Hilfe dieses „echten“ SAR Hotspots möglich die Bedeutung zu erkennen.

Zusammenfassend lässt sich feststellen, dass das unstrukturierte Fingerprint-basierte Ähnlichkeits-Netzwerk leicht mittels inSARa interpretiert werden kann. Das inSARa-Netzwerk wird durch das Teilen des Datensatzes auf eine gut überschaubare Größe reduziert. Wie anhand der Beispiele veranschaulicht, lassen sich schnell bioaktivitätsentscheidende Merkmale erkennen. Obwohl einige heterogene Bereiche im Netzwerk in einigen Fällen auch auf Fehlgruppierungen zurückzuführen sind, wird dieses Risiko durch das Vorfiltern mit den Fingerprints reduziert, wie anhand der gezeigten, gut interpretierbaren Beispiele belegt werden konnte.

21.2. Variante B: Reduktion der MCS-Menge

In Tabelle 21.2 und Tabelle 21.3 sind die Ergebnisse der Analyse zur Variante B des inSARa Hybrid Ansatzes aus Kapitel 14.2 zusammengefasst.

Erwartungsgemäß kann durch die Beschränkung der MCS-Bestimmung durch den Tc-Schwellenwert die Gesamtmenge an einzigartigen MCSs in der MCS-Matrix deutlich reduziert werden (vgl. Tabelle 21.2). Je höher der Tc-Schwellenwert gewählt wird, desto stärker ist die Reduktion. Wie zu erwarten ist die Reduktion bei der Adjazenz-Variante größer als bei der Komponenten-Variante.

Die Reduktion der MCSs in der MCS-Matrix führt dazu, dass die Netzwerk-Komplexität, d.h. die Zahl der MCS-Knoten im Netzwerk, ebenfalls mit steigendem Tc-Schwellenwert abnimmt (bei der Komponenten-Variante schwächer als bei der Adjazenz-Variante). Die Anzahl der Komponenten nimmt dabei erwartungsgemäß zu, da bei steigendem Tc kleine MCSs mit Verknüpfungsfunktion in der MCS-Menge fehlen. Ebenfalls nimmt der Anteil an Molekülen, die nicht im Netzwerk repräsentiert werden, mit steigendem Tc zu. Bei der Adjazenz-Variante ist dies erwartungsgemäß stärker ausgeprägt als bei der Komponenten-Variante, da bei der Komponenten-Variante durch die Nachbarschaftsbeziehungen geringere Tc-Ähnlichkeiten zwischen Molekülen einer Komponente für die Bestimmung des MCS möglich sind.

Vergleicht man die Zunahme des Anteils nicht-repräsentierter Moleküle mit den Ergebnissen aus Tabelle 18.1 in Abschnitt 18.3, so ist die Zunahme geringer als bei der Erhöhung der Mindest-MCS-Größe (bei gleicher oder stärker Reduktion der Netzwerk-Komplexität). Ebenfalls ist festzustellen, dass die Reduktion der MCS-Knoten deutlich stärker ist als durch die Erhöhung des Abbruchkriteriums (vgl. Tabelle 18.3, Abschnitt 18.3). Betrachtet man, welche MCS-Knoten aus der MCS-Menge bzw. aus dem Netzwerk durch die Tc-Restriktion verschwinden, so sind dies vornehmlich MCSs mit einer geringeren MCS-Ähnlichkeit (z.B. MCS-Sim < 0.6). Da diese mehr verknüpfende Funktion haben und meist zu sehr abstrakten, schwer interpretierbaren Molekülgruppierungen führen bzw. ein erhöhtes Risiko für Fehlgruppierungen besteht, ist durch den Verlust dieser MCSs kaum mit einem Informationsverlust in Bezug auf SAR-Analysen zu rechnen.

Tabelle 21.2. Einfluss des Tc-Schwellenwertes auf die Gesamtmenge an einzigartigen MCSs in der MCS-Matrix und die Komplexität und Topologie der resultierenden inSARa-Netzwerke. Abkürzungen: Komp = Berechnung zwischen allen Molekülen aus einer Komponente erlaubt, Adj = MCS-Berechnung nur zwischen Molekülen mit paarweiser Ähnlichkeit oberhalb des Tc-Schwellenwertes.

Target	Tc-Schwellenwert (ECFP4)	0	0.40	0.45	0.50	0.55	0.60	0.70
CB1 (Komp)	MCSs in MCS-Matrix (Min-Größe = 3)	1623	1178	923	876	815	705	419
	MCSs in MCS-Matrix (Min-Größe = 5)	1254	984	822	792	744	658	411
	Wurzel-Knoten im Netzwerk	96	91	88	89	84	80	82
	MCS-Knoten im Netzwerk	665	709	652	632	601	548	320
	Nicht-repräsentierte Moleküle	88	102	114	119	132	154	273
	Repräsentierte Moleküle	1869	1855	1843	1838	1825	1803	1684
	Nicht-repräsentierter Anteil (%)	4.5	5.2	5.8	6.1	6.7	7.9	13.9
	Anzahl an Komponenten	67	61	62	64	62	62	68

CB1 (Adj)	MCSs in MCS-Matrix (Min-Größe = 3)	1623	739	660	614	552	484	325
	MCSs in MCS-Matrix (Min-Größe = 5)	1254	689	625	586	530	471	319
	Wurzel-Knoten im Netzwerk	96	92	98	99	95	96	103
	MCS-Knoten im Netzwerk	665	566	498	464	416	361	209
	Nicht-repräsentierte Moleküle	88	120	127	142	155	173	281
	Repräsentierte Moleküle	1869	1837	1830	1815	1802	1784	1676
	Nicht-repräsentierter Anteil (%)	4.5	6.1	6.5	7.3	7.9	8.8	14.4
	Anzahl an Komponenten	67	64	68	67	65	73	90
P38 (Komp)	MCSs in MCS-Matrix (Min-Größe = 3)	2521	2117	1695	1558	1333	1189	771
	MCSs in MCS-Matrix (Min-Größe = 5)	1980	1722	1477	1380	1236	1120	751
	Wurzel-Knoten im Netzwerk	119	118	117	115	107	105	107
	MCS-Knoten im Netzwerk	1106	1197	1128	1093	1010	955	615
	Nicht-repräsentierte Moleküle	48	56	64	71	90	109	224
	Repräsentierte Moleküle	2398	2390	2382	2375	2356	2337	2222
	Nicht-repräsentierter Anteil (%)	2.0	2.3	2.6	2.9	3.7	4.5	9.2
	Anzahl an Komponenten	67	55	56	57	54	55	82
P38 (Adj)	MCSs in MCS-Matrix (Min-Größe = 3)	2521	1312	1191	1107	987	883	607
	MCSs in MCS-Matrix (Min-Größe = 5)	1980	1224	1134	1063	956	863	599
	Wurzel-Knoten im Netzwerk	119	114	116	121	127	131	158
	MCS-Knoten im Netzwerk	1106	1011	936	872	778	752	428
	Nicht-repräsentierte Moleküle	48	76	81	89	103	118	254
	Repräsentierte Moleküle	2398	2370	2365	2357	2343	2328	2192
	Nicht-repräsentierter Anteil (%)	2.0	3.1	3.3	3.6	4.2	4.8	10.4
	Anzahl an Komponenten	67	58	57	69	79	87	124
THR (Komp)	MCSs in MCS-Matrix (Min-Größe = 3)	3963	3632	3247	2849	2673	2353	1390
	MCSs in MCS-Matrix (Min-Größe = 5)	3484	3212	2929	2622	2479	2221	1371
	Wurzel-Knoten im Netzwerk	105	107	107	111	107	101	114
	MCS-Knoten im Netzwerk	2379	2384	2326	2179	2076	1940	1222
	Nicht-repräsentierte Moleküle	57	57	59	66	75	101	204
	Repräsentierte Moleküle	2795	2795	2793	2786	2777	2751	2648
	Nicht-repräsentierter Anteil (%)	2.0	2.0	2.1	2.3	2.6	3.5	3.2
	Anzahl an Komponenten	44	43	39	46	50	52	79
THR (Adj)	MCSs in MCS-Matrix (Min-Größe = 3)	3963	2183	1973	1766	1578	1369	924
	MCSs in MCS-Matrix (Min-Größe = 5)	3484	2085	1894	1703	1528	1333	920
	Wurzel-Knoten im Netzwerk	105	116	121	142	147	178	200
	MCS-Knoten im Netzwerk	2379	1793	1626	1434	1266	1054	684
	Nicht-repräsentierte Moleküle	57	65	73	80	100	134	240
	Repräsentierte Moleküle	2795	2787	2779	2772	2752	2718	2612
	Nicht-repräsentierter Anteil (%)	2.0	2.3	2.6	2.8	3.5	4.7	8.4
	Anzahl an Komponenten	44	60	69	82	84	117	155

Betrachtet man zusätzlich zur Veränderung der Netzwerk-Komplexität und -Topologie wie sich die Homogenität der Netzwerke bzw. Anteil an SAR-(Dis-)Kontinuität verändert (vgl. Tabelle 21.3), so lässt sich Folgendes beobachten: Der Anteil an kontinuierlicher und diskontinuierlicher SAR-Information bleibt bei Erhöhung des Tc-Schwellenwertes annährend konstant. Vereinzelt ist erwartungsgemäß eine Verschiebung in Richtung SAR-Kontinuität (Erhöhung des SARdisco) zu beobachten. Dies ist damit zu erklären, dass „künstliche“ SAR-Heterogenität oder Diskoninuität („falschpositive“ ACs/SAR Hotspots) aufgrund von molekularen Fehlgruppierungen bedingt durch den Abstraktionslevel der RGs durch den Fingerprint-basierten Vorfilter reduziert werden. Dies drückt sich auch in der Abnahme des Medians des MAXAD aus. Der Median des MAD bleibt annähernd konstant.

Tabelle 21.3. Einfluss des Tc-Schwellenwertes auf die Reinheit der inSARa-Netzwerke bzw. SAR-(Dis-)Kontinuität in den Netzwerken. Abkürzungen: Komp = Berechnung zwischen allen Molekülen aus einer Komponente erlaubt, Adj = MCS-Berechnung nur zwischen Molekülen mit paarweiser Ähnlichkeit oberhalb des Tc-Schwellenwertes. MAD = Median der absoluten Bioaktivitäts-Abweichung (pK_i) am MCS-Knoten X, angegeben wird der Median aller Knoten mit MCS-Sim > 0.6; MAXAD = maximale Bioaktivitäts-Abweichung (pK_i) vom Median am MCS-Knoten X, angegeben wird (wie bei MAD) der Median aller Knoten mit MCS-Sim > 0.6.

Variante			**Komp**			**Adj**		
Target	**Tc-Schwellenwert (ECFP4)**	**0**	**0.4**	**0.55**	**0.7**	**0.4**	**0.55**	**0.70**
CB1	Median(MAD)	0.32	0.34	0.33	0.32	0.34	0.32	0.31
	Median(MAXAD)	2.36	1.52	1.53	1.55	1.51	1.46	1.47
	SARdisco Score	0.70	0.69	0.68	0.70	0.69	0.70	0.70
P38	Median(MAD)	0.25	0.26	0.26	0.25	0.26	0.25	0.25
	Median(MAXAD)	2.23	1.03	1.05	1.03	1.03	0.97	1.0
	SARdisco Score	0.79	0.79	0.79	0.81	0.79	0.80	0.82
THR	Median(MAD)	0.23	0.35	0.35	0.33	0.33	0.30	0.28
	Median(MAXAD)	2.68	1.56	1.55	1.44	1.41	1.24	1.34
	SARdisco Score	0.61	0.61	0.62	0.64	0.64	0.68	0.71

21.3. Zusammenfassung

Zusammenfassend lässt sich feststellen, dass die Kombination des inSARa-Konzeptes mit Fingerprint-Ähnlichkeit für die SAR-Analyse sehr vorteilhaft sein kann. Der inSARa Hybrid Ansatz stellt v.a. für sehr große Datensätze eine weitere Möglichkeit dar (zusätzlich zu den in Kapitel 18 diskutierten Optimierungsparametern), die Netzwerk-Komplexität zu reduzieren und somit die Interpretierbarkeit der Netzwerke weiter zu verbessern. Durch Vorclustern von Molekülen und anschließende Clusteranalyse können auch sehr große Datensätze (wie am Beispiel des Thrombin-Datensatzes gezeigt) mittels inSARa ohne Schwierigkeiten analysiert werden. Ein weiterer Vorteil für die SAR-Interpretation ist, dass durch das Vorfiltern mit dem ECFP4-Fingerprint molekulare Fehlgruppierungen an MCS-Knoten bzw. künstliche ACs/SAR Hotspots (z.B. aufgrund der RG-Abstraktion) reduziert werden können. Bei der Wahl des Tc-Schwellenwertes bzw. der Entscheidung zwischen Adjazenz- und Komponenten-Variante ist auch hier (analog zu Kapitel 18.3) ebenfalls ein Kompromiss aus potentiellem Informationsverlust und Zunahme an Interpretierbarkeit zu suchen. Je nach Fragestellung und Datensatz lassen sich die inSARa-Netzwerke über die aufgezeigten Parameter individuell optimieren.

22. Ergebnisse und Diskussion: inSARa^{auto} und SARdisco

22.1. Ergebnisse der automatisierten SAR-Analyse verschiedener Datensätze aus der BindingDB (inSARa^{auto})

Abbildung 22.1 zeigt einen Vergleich verschiedener großer BindingDB-Datensätze bezüglich enthaltener SAR-Information basierend auf der automatisierten Auswertung der zugehörigen inSARa-Netzwerke mittels inSARa^{auto}. Wie zu erwarten weisen die Datensätze deutliche Unterschiede im Hinblick auf kontinuierliche (bioaktivitätserhaltende Modifikationen) und diskontinuierliche (sprunghafte SARs und SAR Hotspots) SAR-Information auf.

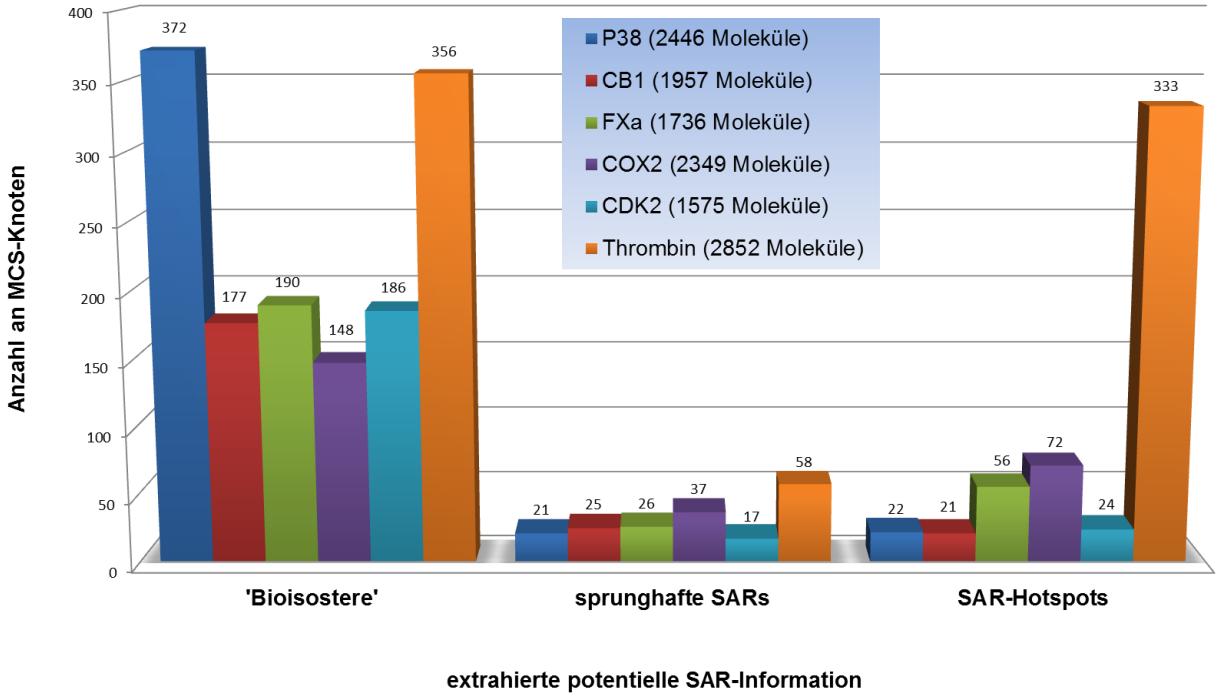


Abbildung 22.1. Automatisierte Analyse verschiedener BindingDB-Datensätze bezüglich enthaltener SAR-Information mittels inSARa^{auto}. Anzahl der MCS-Knoten in den inSARa-Netzwerken, die potentiell SAR-Information eines bestimmten Typs (bioaktivitätserhaltende Modifikation, sprunghafte SARs oder SAR Hotspots) enthalten.

So enthält der P38- und COX2-Datensatz eine vergleichbare Anzahl an Molekülen, das COX2-inSARa-Netzwerk enthält jedoch nur halb so viele „Bioisosterie“-MCS-Knoten wie das P38-Netzwerk. Stattdessen weist es fast die doppelte Menge AC-MCS-Knoten und fast die dreifache Menge an SAR Hotspot-MCS-Knoten auf. Während das P38-Netzwerk hauptsächlich kontinuierliche SAR-Information enthält, weist das COX2-Netzwerk einen ebenfalls hohen Anteil diskontinuierlicher SAR-Information auf. Dies ist auch konsistent mit den entsprechender Fingerprint-basierten globalen Kontinuitäts- und Diskontinuitäts-Werten für diese Datensätze (vgl. Tabelle 26.2 im Anhang). Vergleicht man den Thrombin-Datensatz mit dem P38-Datensatz, so weisen die entsprechenden Netzwerke eine vergleichbare Menge an „Bioisosterie“-MCS-Knoten auf. Jedoch ist das Besondere beim Thrombin-

Datensatz (auch im Vergleich zu allen anderen Datensätzen), dass eine fast ebenso große Anzahl an SAR Hotspot-MCS-Knoten und eine hohe Anzahl AC-MCS-Knoten im Netzwerk enthalten sind. Der Thrombin-Datensatz enthält somit einen hohen Anteil kontinuierlicher als auch diskontinuierlicher SAR-Information. Dies spiegelt sich ebenfalls in dem globalen FP-basierten (Dis-)Kontinuitäts-Wert wider (vgl. Tabelle 26.2). Vergleichbar bezüglich der Anzahl an MCS-Knoten mit diskontinuierlicher Information mit dem P38-Datensatz ist der CB1- und CDK2-Datensatz. Beide Datensätze zeigen ein ähnliches Verhalten. Im Vergleich zu P38 weisen die Netzwerke etwa nur halb so viele „Bioisosterie"-MCS-Knoten auf. Das FXa-Netzwerk hingegen ist bezüglich der Anzahl an MCS-Knoten mit kontinuierlicher SAR-Information mit CDK2 und CB1 vergleichbar, im Hinblick auf MCS-Knoten mit diskontinuierlicher SAR-Information (v.a. SAR Hotspots) ähnelt es mehr dem COX2-Netzwerk. FXa nimmt eine Zwischenstellung zwischen CDK2/CB1 und COX2 ein. Der globale FP-basierte (Dis-)Kontinuitäts-Wert für FXa (vgl. Tabelle 26.2) hätte einen geringeren kontinuierlichen und höheren diskontinuierlichen Anteil erwarten lassen. Auch hätte man beispielsweise für COX2 und CB1 einen höheren kontinuierlichen Anteil im Vergleich zu P38 oder Thrombin erwartet.

Bezüglich MCS-Knoten, die SAR-Kontinuität repräsentieren, im Verhältnis zu MCS-Knoten, die diskontinuierliches SAR-Verhalten repräsentieren, lässt sich bei den analysierten Datensätzen folgende Reihenfolge (abnehmende SAR-Kontinuität) feststellen: P38 > CDK2 ≈ CB1 > FXa > COX2 > Thrombin. Auf Basis FP-basierter SARI-Werte resultiert die folgende Reihenfolge (abnehmender SARI-Wert = abnehmende SAR-Kontinuität): CB1 ≈ P38 >CDK2 > COX2 > FXa > Thrombin. Es ist zwar in einigen Fällen ein ähnlicher Trend (z.B. P38 und Thrombin) erkennbar, es sind jedoch auch einige Unterschiede zwischen FP- und inSARa-basierten Analyse (z.B. CB1, COX2 oder FXa) festzustellen.

22.2. Diskussion: inSARa^{auto}

inSARa^{auto} ermöglicht eine sehr schnelle SAR-Analyse, d.h. es lässt sich schnell erkennen, welche SAR-Information potentiell in inSARa-Netzwerken verborgen ist. Die entsprechenden Knoten können entweder sortiert nach zu erwartender SAR-Information aus dem Netzwerk extrahiert oder im ursprünglichen Netzwerk hervorgehoben werden. Durch die regelbasierte Automatisierung der SAR-Informations-Erkennung wird die Datensatz-Analyse objektiver und die wichtige Information wird schneller in den Netzwerken gefunden. Zudem lassen sich verschiedene Datensätze so sehr einfach bezüglich enthaltener Information vergleichen. Für FP-basierte SAR-Analyse sind zwar quantitative Maßzahlen zur Datensatz-Charakterisierungen wie numerische (Dis-)Kontinuitäts- und SARI-Werte (vgl. Abschnitt 2.4.2) und automatisierte Erkennung von „SAR Pathways" beschrieben (vgl. Abschnitt 2.6.4), eine automatisierte SAR-Informations-Erkennung der Form wie sie inSARa^{auto} ermöglicht, ist bisher nicht beschrieben. Im Vergleich zu FP-basierten (Dis-)Kontinuitäts- und SARI-Werten liegen die Stärken der inSARa-basierten Datensatz-Auswertung, wie schon mehrfach betont, in der Intuitivität und direkten Interpretierbarkeit. Im Gegensatz zu FP-basierter Analyse können eventuelle Fehl-Klassifikationen (z.B. aufgrund von suboptimalen Grenzwerten oder mangelhafter RG-Codierung) durch visuelle Inspektion der entsprechenden MCS-Knoten schnell erkannt werden. FP-basierte SAR-(Dis-)Kontinuität, die durch mangelhafte molekulare Repräsentation bedingt ist, ist jedoch schwer zu prüfen. Eine Anpassung der für inSARa^{auto} verwendeten Grenzwerte ist jederzeit möglich.

22.3. Ergebnisse der globalen automatisierten Charakterisierung verschiedener BindingDB-Datensätze bezüglich SAR-(Dis)Kontinuität (SARdisco)

In Abbildung 22.2 sind die Ergebnisse der quantitativen Analyse der 140 Datensätze aus der BindingDB mittels inSARa-Netzwerk-basiertem SARdisco und Fingerprint-basiertem SAR-Index (auf Basis von MACCS Keys) zusammengefasst.

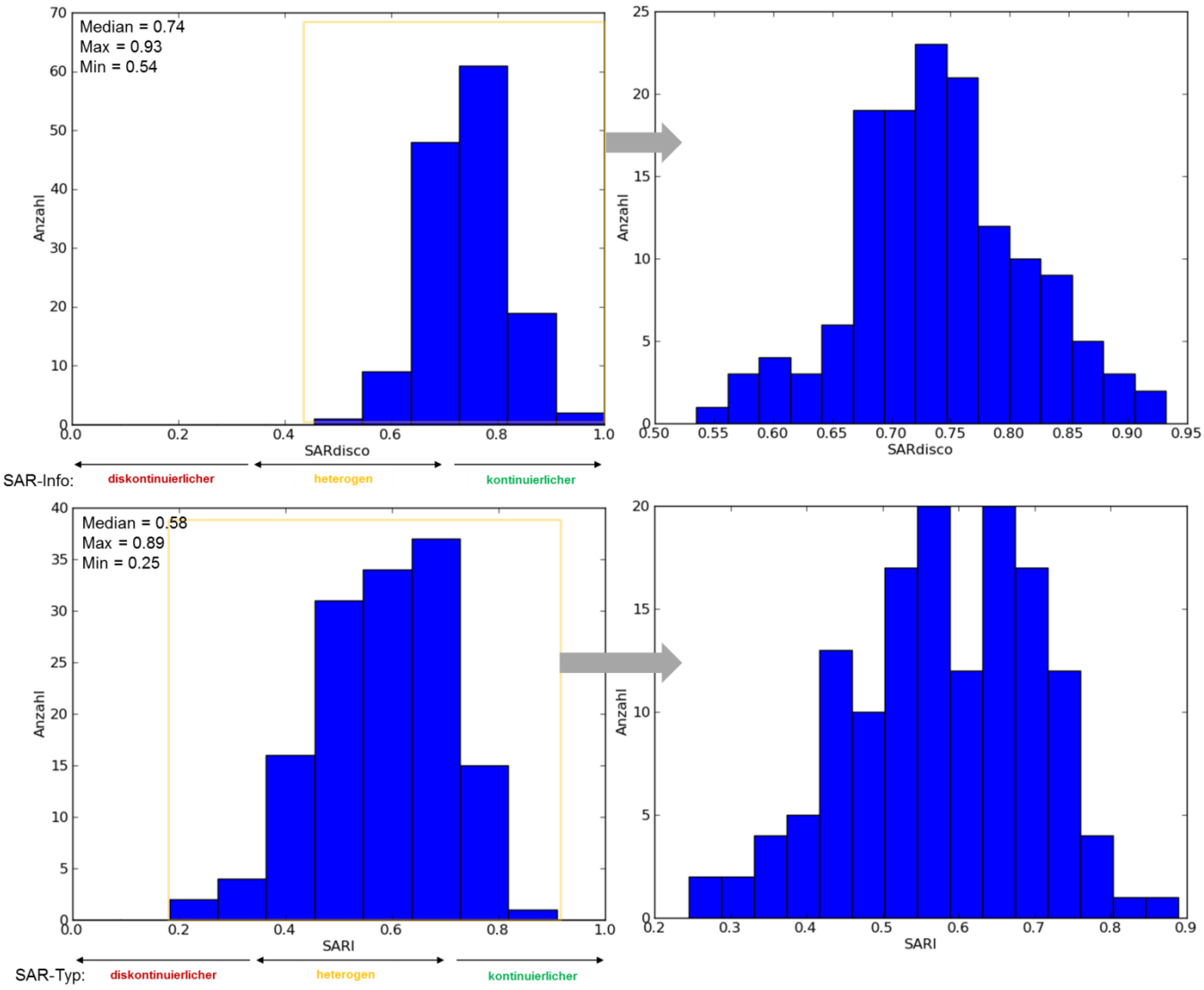


Abbildung 22.2. Vergleich der Häufigkeitsverteilung der SARdisco (oben) und SAR-Index Werte (unten) für die 140 BindingDB-Datensätze.

Anhand der SARdisco Verteilung (Abbildung 22.2, oben) ist zu erkennen, dass die inSARa-Netzwerke der meisten Datensätze einen hohen bis sehr hohen Anteil an kontinuierlicher SAR-Information (SARdisco > 0.7) enthalten. In keinem der analysierten Netzwerke repräsentiert die Mehrheit an MCS-Knoten diskontinuierliche SAR-Information (SARdisco < 0.5). Einige wenige Netzwerke sind durch stärkere Heterogenität bezüglich repräsentierter SAR-Information (SARdisco < 0.65) bzw. durch Repräsentation fast ausschließlich von SAR-

Kontinuität (SARdisco > 0.85) geprägt. Für die sechs Beispiel-Datensätze aus Tabelle 11.1 ergibt mittels SARdisco dieselbe Reihenfolge abnehmender SAR-Kontinuität, die sich auch auf Basis der inSARa^{auto}-Analyse (vgl. Abschnitt 22.1) ergeben hat: P38 > CDK2 > CB1 > FXa > COX2 > THR. Die zusätzliche Berücksichtigung heterogener MCS-Knoten und Aufhebung der Mindestaktivitätsschwelle MCS-Knoten, die homogen bezüglich der Aktivitätsverteilung sind, hat in diesen Fällen keine signifikante Auswirkung auf das Ergebnis gezeigt.

Vergleicht man diese Ergebnisse mit der SAR-Index Verteilung (Abbildung 22.2, unten) für diese Datensätze, so werden die meisten Datensätze als heterogener SAR-Typ charakterisiert (SARI: etwa 0.4 bis 0.6). Zusätzlich lässt sich ein ebenfalls hoher Anteil an Datensätzen als eher kontinuierlicher SAR-Typ (SARI: etwa 0.6 bis 0.8) einordnen. Ein kleiner Anteil an Datensätzen lässt sich hingegen als eher diskontinuierlicher SAR-Typ (SARI: etwa 0.2 bis 0.4) charakterisieren.

In Abbildung 22.3 sind die jeweiligen SARdisco und SARI Werte für alle Datensätze gegeneinander aufgetragen. Der eher geringe Korrelationskoeffizienten ρ_S von 0.5 zeigt, dass beide Maßzahlen zur globalen Charakterisierung von SAR-(Dis-)Kontinuität nur schwach korrelieren und anhand des auf Basis von Fingerprint-Ähnlichkeit berechneten SAR-Index beispielsweise nur schwer ein Rückschluss auf den SARdisco Wert bzw. die Homogenität oder Heterogenität der resultierenden inSARa-Netzwerke möglich ist.

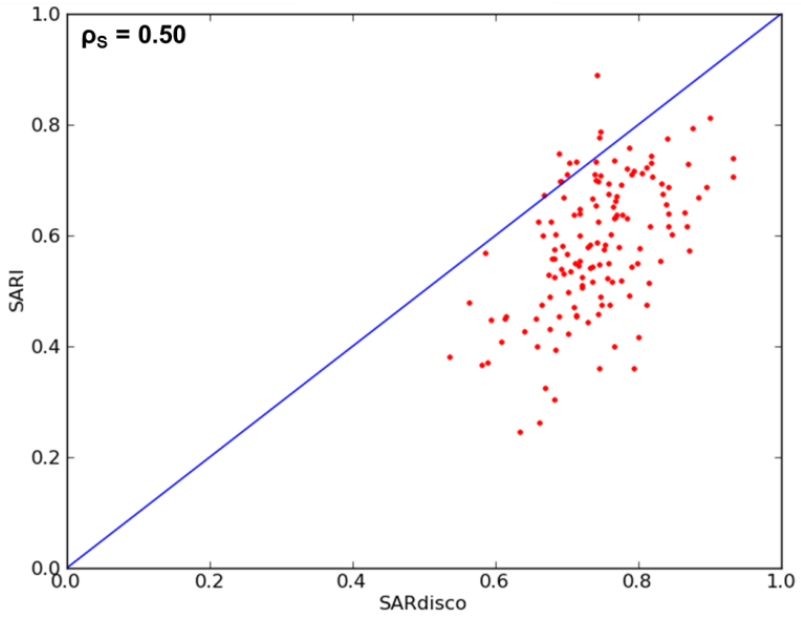


Abbildung 22.3. Korrelation (Spearman-Rang-Korrelationskoeffizient ρ_S) zwischen dem dem inSARa-basierten SARdisco und Fingerprint-basierten SAR-Index

22.4. Diskussion: SARdisco

SAR-Kontinuität und Diskontinuität sind stark von molekularen Repräsentation und dem verwendeten Ähnlichkeitsmaß abhängig (vgl. Kapitel 2.4). Bei schlechter molekularer

Repräsentation ist ein hoher Anteil an SAR-Heterogenität oder Diskontinuität zu erwarten. Gute molekulare Repräsentation sollte in Netzwerken mit hoher SAR-Kontinuität bzw. Datensätzen mit hohem SAR-Index oder Netzwerken mit hohem SARdisco resultieren.

Auf Grundlage von inSARa-Netzwerk-basierter Aktivitätslandschafts-Analyse lässt sich feststellen, dass der Anteil zerklüfteter Aktivitätslandschaft nach MAGGIORA („Bryce-Canyon-Metapher“, vgl. Abschnitt 2.4.1) im Vergleich zu homogener Aktivitätslandschaft („geschwungene Hügel“, vgl. Abschnitt 2.4.1) in der Mehrheit an Datensätzen eher gering ausfällt. Fingerprint-basierte SARI-Analysen hingegen zeigen im Vergleich dazu oftmals eine Verschiebung von homogener zu heterogenerer Aktivitätslandschaft an. Eine wahrscheinliche Ursache hierfür ist die weniger adäquate molekulare Repräsentation durch den zugrunde liegenden Fingerprint (MACCS Keys). Dies ist auch im Einklang mit den Ergebnissen der *k*NN-Regression aus Kapitel 19 (Unterlegenheit der MACCS Keys gegenüber inSARa). Bei dieser Analyse impliziert ein geringerer absoluter oder relativer Fehler ebenfalls mehr SAR-Homogenität. Die Aussagekraft eines geringen oder mittleren SARI-Wertes ist für die SAR-Analysen somit deutlich eingeschränkt. Kleine SARdisco-Werte zeigen auch nur an, dass viele heterogene MCS-Knoten im inSARa-Netzwerk zu finden sind. Eine suboptimale RG-Codierung ist neben „echter“ SAR-Diskontinuität eine weitere mögliche Ursache. Jedoch sind auch hohe SARI- oder SARdisco-Werte (= hohe SAR-Kontinuität) ebenfalls mit Vorsicht zu betrachten. So kann künstliche SAR-Kontinuität beispielsweise auch aus einer ungleichen Aktivitätsverteilung in Datensätzen (z.B. nur schwach oder mittel aktive Moleküle) resultieren.

In Bezug auf sprunghafte SARs lässt sich bei der inSARa-Analyse der Datensätze feststellen, dass in der Regel in jedem Datensatz Beispiele hierfür identifiziert werden können. Dies bestätigt die These von MAGGIORA, dass sprunghafte SARs ein fester Bestandteil der Aktivitätslandschaft sind (vgl. Kapitel 2.5.1). Zahlenmäßig in Bezug auf die gesamte Aktivitätslandschaft bzw. die Gesamtgröße des Datensatzes gesehen, machen ACs jedoch nur einen vergleichsweise kleinen Anteil aus. Häufiger sind hingegen SAR Hotspots in Datensätzen zu finden. In Kapitel 2.5.1 wurde beschrieben, dass zahlreiche Analysen der letzten Jahre zeigen, dass ACs sehr häufig in Datensätzen vorkommen. Hierbei ist zu berücksichtigen, dass es sich hierbei zumeist um Analysen auf Basis von paarweisen Molekülvergleichen handelt. Bei paarweiser AC-Definition (ohne Berücksichtigung weiterer Nachbarmoleküle) werden auch Molekülpaare erfasst, die eigentlich kein klassisches AC bilden, sondern Teil eines SAR Hotspots sind. Die klare Unterscheidung zwischen AC und SAR Hotspot ist mittels Substruktur-basierter Ansätze wie inSARa deutlich einfacher möglich.

Der SARdisco ermöglicht (analog zum SAR-Index) inSARa-Netzwerke bezüglich vorhandener SAR-(Dis-)Kontinuität global zu charakterisieren. Diese Maßzahl kann beispielsweise zur Optimierung der Netzwerke verwendet werden. Im Unterschied zum SAR-Index ist die Berechnung deutlich einfacher und es sind keine Kalibrierdatensätze bzw. eine aufwändige Z-Score-basierte Reskalierung der Rohwerte notwendig. Beide Maßzahlen sind erwartungsgemäß (vgl. Ergebnisse aus Abschnitt 18.2) aufgrund der unterschiedlichen molekularen Repräsentation und Ähnlichkeitserfassung, sowie konzeptionell etwas unterschiedlichen Berechnung nicht direkt miteinander korreliert.

23. Ergebnisse und Diskussion: inSARa-Netzwerk-Vergleich

23.1. Ergebnisse

Bei der Analyse aus Kapitel 17 und den nachfolgend vorgestellten Ergebnissen war als eigentliches Ziel ein Vergleich der für SAR-Analysen optimierten inSARa-Netzwerke verschiedener Targets definiert und nicht die Entwicklung einer optimierten Chemogenomik-Methode. Die nachfolgend gezeigten Ergebnisse der inSARa/RG-MCS-basierte Analyse sollen daher vorwiegend veranschaulichen, dass interpretierbare, sinnvolle Beziehungen aus dem Netzwerk-Vergleich resultieren, um abermals die Sinnhaftigkeit der dem in inSARa-Ansatz zugrunde liegenden gewählten molekularen Repräsentation und des Ähnlichkeitsvergleiches zu unterstreichen.

Da durch Kombination verschiedener Methoden ein umfassenderer Einblick in die komplexen polypharmakologischen Beziehungen des Liganden- und Target-Universums erhalten werden kann, soll zudem untersucht werden, ob dieser neuartige RG-MCS-basierte Ansatz das Potential der sinnvollen Ergänzung verfügbarer Chemogenomik-Ansätze besitzt. Hierzu werden Vor- und Nachteile bzw. Grenzen des Verfahren bzw. der Analyse aufgezeigt.

23.1.1. Analyse verschiedener Einflussgrößen

Bei Chemogenomik-Ansätzen sind die Ergebnisse von zahlreichen Einflussfaktoren (z.B. Datengrundlage[330], molekulare Repräsentation/Methode[346], Mindest-Aktivitätsschwelle[342]) abhängig. Auch bei dieser Analyse werden die gefundenen Ähnlichkeits-Beziehungen von vielen Faktoren beeinflusst: Anzahl der Liganden im Datensatz und deren strukturelle Diversität, Art der RG-Codierung der Moleküle, Wahl des Mindest-Aktivitäts-Schwellenwert, Größenunterschiede der Datensätze bzw. der MCS-Mengen und Art der Berechnung der Target-Ähnlichkeiten. Die Analyse kann somit niemals den Anspruch einer vollständigen Beschreibung (poly-)pharmakologischer Beziehungen haben, sondern ist als Ergänzung der der ebenfalls begrenzten Erkenntnisse zu sehen, die mit anderen Methoden (vgl. Abschnitt 8.2) gewonnen werden. Im Folgenden werden von den genannten einige der derjenigen Einflussfaktoren (Größenunterschiede in MCS-Mengen und Art der Ähnlichkeitsberechnung) näher untersucht, die spezifisch für diesen Ansatz sind.

a) Einfluss der Berücksichtigung von Substruktur-Beziehungen und Unterschieden in der Datensatzgröße bzw. MCS-Menge („Selbstähnlichkeitstest“)

Die Ergebnisse des in Abschnitt 17.4 beschriebenen „Selbstähnlichkeitstestes“ zur Analyse des Einflusses der Größenunterschiede der zu vergleichenden MCS-Mengen, sowie des Einflusses der Berücksichtigung von Substruktur-Beziehungen bei der Bestimmung der gemeinsamen MCSs von zwei Targets (vgl. Abschnitt 17.2 → Schritt 2) sind in Abbildung 23.1 (für Thrombin) bzw. Abbildung 26.3 und Abbildung 26.4 in Abschnitt 26.6 im Anhang (für COX2 und P38) zusammengefasst.

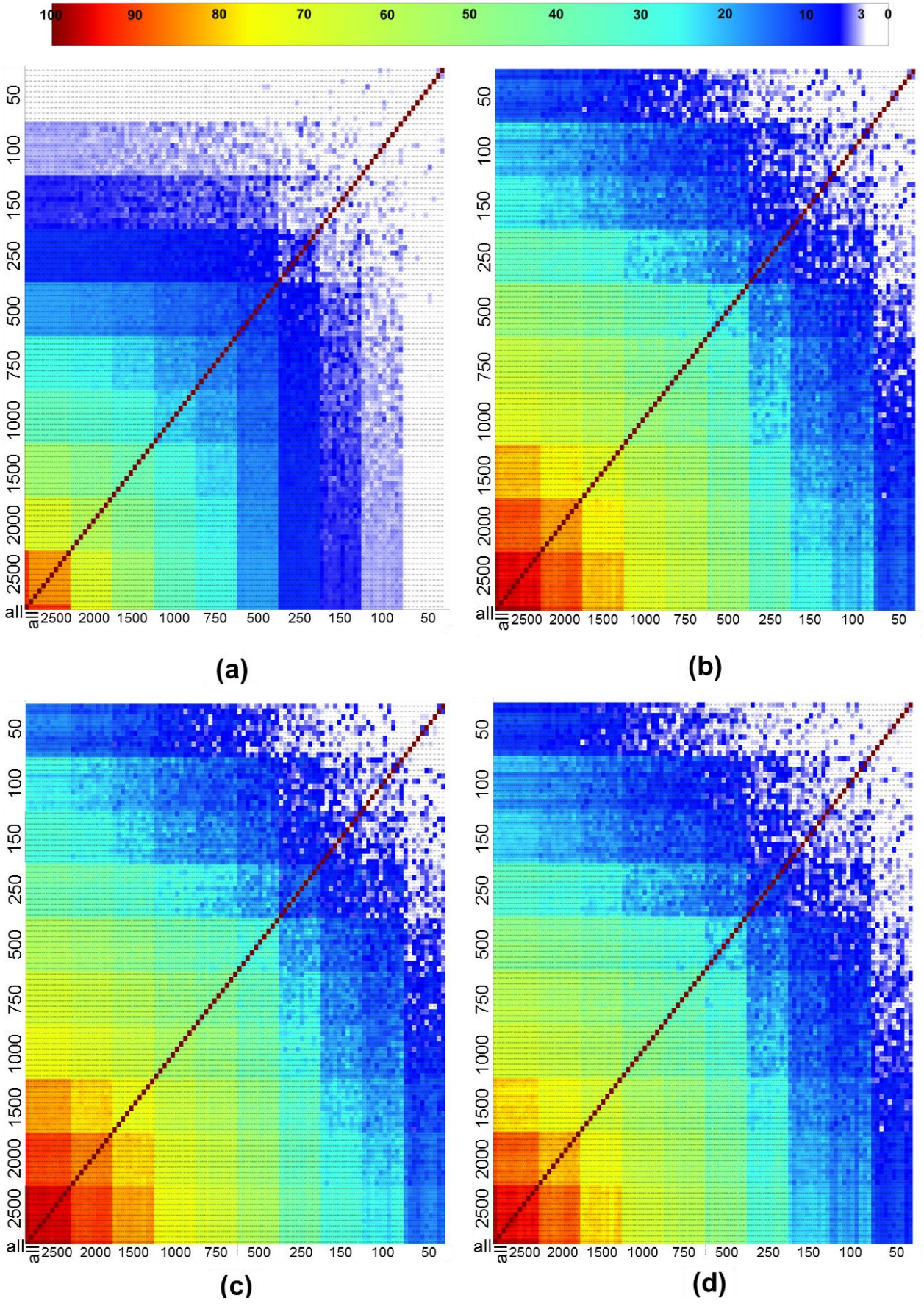


Abbildung 23.1. Ergebnisse des „Selbstähnlichkeitstestes“ am Beispiel des Thrombin-Datensatzes: (a) *ohne*, (b) bis (d) *mit* Berücksichtigung von Substruktur-Beziehungen. Einfluss der Gewichtung: (a) und (b) *gewichteter* inSARa-TSim, (c) *ungewichteter* inSARa-TSim, (d) *quadrierter Tc*.

Bei der Beschriftung der Heatmaps in Abbildung 23.1, Abbildung 26.3 und Abbildung 26.4 stellt „all“ jeweils die gesamte MCS-Menge des entsprechenden Targets dar. Die Zahlen geben jeweils die Größe der Stichprobe an, die aus „all“ zufällig gezogen wurde (jeweils 10 Wiederholungen pro Stichprobengröße). So bedeutet „500“ in Abbildung 23.1beispielsweise, dass es sich hierbei um 500 zufällig gezogene MCSs aus der Gesamt-MCS-Menge von Thrombin (insgesamt 2732 MCSs) handelt.

Vergleicht man die Heatmaps (a) bis (d) in Abbildung 23.1, so ist mit zunehmendem MCS-Größenunterschied (z.B. „all“ in Vergleich zu „2500“, „500“ und „50“) eine Farbverschiebung von Rot nach Blau zu beobachten. Dies veranschaulicht, dass Unterschiede in der Größe der zu vergleichenden MCS-Mengen einen deutlichen Einfluss auf die berechnete Ähnlichkeit haben. Wie zu erwarten, ist die berechnete Ähnlichkeit umso geringer, je mehr sich die beiden zu vergleichenden MCS-Mengen in der Größe unterscheiden.

Vergleicht man den Einfluss der Berücksichtigung von Substruktur-Beziehungen beispielhaft für Thrombin (Heatmap (a) und (b) inAbbildung 23.1), so ist anhand der Farben klar zu erkennen, dass eine deutlich höhere Ähnlichkeit zwischen den verschiedenen MCS-Mengen unter Berücksichtigung von Substruktur-Beziehungen (b) im Vergleich zu (a) (ohne Berücksichtigung von Substruktur-Beziehungen) gefunden werden kann. Da die verschiedenen MCS-Mengen beim „Selbstähnlichkeitstest“ von demselben Target stammen, sollte sich eine möglichst hohe Ähnlichkeit zwischen zwei zu vergleichenden MCS-Mengen ergeben. Anhand des Vergleiches der gesamten („all“) mit der kleinsten MCS-Stichprobe („50“ in der untersten Reihe der Heatmaps) kann man sehen, dass aufgrund der Berücksichtigung von Substruktur-Beziehungen (b) eine deutliche Ähnlichkeit oberhalb des für die Analysen definierten Ähnlichkeitsschwellen-Wertes von 3.0 gefunden wird. Ohne die Berücksichtigung von Substruktur-Beziehungen (a) ist für kleinere MCS-Mengen keine oder nur noch eine sehr geringe Ähnlichkeit festzustellen (vgl. „50“er, „100“er und „150“er Stichproben). Ein ähnliches Verhalten zeigt sich auch bei den zwei weiteren Datensätzen (COX2 und P38, vgl. Heatmap (a) und (b) in Abbildung 26.3 und Abbildung 26.4 im Anhang). Die Berücksichtigung von Substruktur-Beziehungen ist somit vorteilhaft für das Erkennen von Ähnlichkeit zwischen MCS-Mengen, die größere Größenunterschiede aufweisen. Die hierdurch bedingten Verzerrungen bei der Bestimmung der Ähnlichkeit lassen sich zu einem gewissen Maß dadurch ausgleichen. Ebenso ist hierdurch eine bessere Abgrenzung zwischen ähnlichen (höherer Gewichtung durch zusätzliche Substruktur-MCSs in der Menge gemeinsamer MCSs) und unähnlichen MCS-Mengen möglich.

Vergleicht man die verschiedenen Gewichtungsmöglichkeiten bei der Berechnung der Ähnlichkeit, so ist anhand des Farbvergleiches der Heatmaps (b) bis (d) in Abbildung 23.1 festzustellen, dass ohne Gewichtung mit der MCS-Größe (c) die größte Ähnlichkeit gefunden wird. Berücksichtigt man die MCS-Größe bei der Gewichtung, so ist wird mittels des gewichteten inSARa-TSim-Score (b) eine höhere Ähnlichkeit als mittels quadriertem Tc (d) gefunden. Da die MCS-Größe wichtig für die Bedeutsamkeit einer Ähnlichkeitsbeziehung ist, wurde für alle durchgeführten Analysen daher der gewichtete inSARa-TSim zur Ähnlichkeitsberechnung verwendet.

b) Erkennen von kleinen gleichen Submengen in großen MCS-Mengen („Wiederfindungstest“)

Der in Abschnitt 17.4 beschriebene „Wiederfindungstestes“ dient zur Analyse, ob beim Vergleich von zwei großen MCS-Mengen auch auf Basis einer kleinen gemeinsamen Submenge eine Ähnlichkeit gefunden werden kann. Die zugehörigen Ergebnisse sind in Abbildung 23.2 und Abbildung 23.3 dargestellt (Verwendung des gewichteten inSARa-TSim zur Ähnlichkeits-Berechnung). Der Unterschied zwischen Abbildung 23.2 und Abbildung 23.3 liegt in der (Nicht-)Berücksichtigung von Substruktur-Beziehungen bei der Bestimmung der gemeinsamen MCS-Menge zweier Targets (vgl. Abschnitt 17.2 → Schritt 2).

Die Beschriftung der Mini-Heatmaps (I) bis (III) in den beiden Abbildungen ist wie nachfolgend beschrieben zu verstehen:

a) Die Abkürzungen in der Beschriftung am rechten Rand („all_Target“) gibt jeweils an, dass es sich um die Gesamt-MCS-Menge eines der drei analysierten Targets handelt. „all_P38“ enspricht der Gesamt-MCS-Menge von P38, „all_COX2“ ist die Gesamt-MCS-Menge von COX2 und „all_THR“ die entsprechende Menge von Thrombin.

b) In der unteren Beschriftung haben diese Abkürzungen dieselbe Bedeutung. Die restlichen Abkürzungen (z.B. „9_50_P38“ in Heatmap (I)) enthalten folgende Information: Die erste Zahl gibt jeweils die Stichprobennummer an (im Beispiel: „Nr. 9“). Die zweite Zahl („50“) gibt in Verbindung mit der römischen Nummer der Heatmap (I=Thrombin, II=COX2, III=P38) an, dass 50 zufällig gezogene MCSs des jeweiligen Targets (im Beispiel: Thrombin) der Gesamt-MCS-Menge eines anderen Targets (dritte Angabe in Abkürzung, im Beispiel demzufolge P38) hinzugefügt worden sind. Das heißt, es wurden bei jeder Heatmap jeweils 10 verschiedene Zufalls-Stichproben, die jeweils 50 MCSs aus der Gesamt-Menge eines bestimmten Targets enthalten, gezogen und jeweils der Gesamt-Menge der anderen beiden Targets hinzugefügt. Diese MCS-Mengen werden dann jeweils mit der Gesamt-Menge aller drei Targets verglichen. Zur Kontrolle der Grund-Ähnlichkeit werden zusätzlich auch die Gesamt-MCS-Mengen aller drei Targets miteinander verglichen.

Unter Berücksichtigung von Substruktur-Beziehungen (Abbildung 23.2) ist für alle drei Targets festzustellen, dass auf Basis von 50 zufällig ausgewählten MCSs eine Ähnlichkeit zwischen der Gesamtmenge (Blaufärbung in Heatmap (I) bei „all_THR“, in Heatmap (II) bei „all_COX2“, in Heatmap (III) bei „all_P38“) und der jeweiligen Submenge gefunden werden kann, die oberhalb des verwendeten Ähnlichkeits-Schwellenwertes von 3.0 liegt. Beim paarweisen Vergleich der Gesamt-Mengen der drei Targets, die keine hinzugefügten Submengen eines anderen Targets enthalten, wird dieser Ähnlichkeitswert erwartungsgemäß nicht erreicht (Ausnahme: P38 und COX2).

Ohne Berücksichtigung von Substruktur-Beziehungen (Abbildung 23.3) ist z.B. in Heatmap (I) bei „all_THR“ anstelle des blauen Farbbandes aus Abbildung 23.2 eine Weißfärbung vorzufinden, die andeutet, dass die Ähnlichkeit zwischen den jeweiligen MCS-Mengen nur sehr gering ist. Bei COX2 (II) und P38 (III) wird zwar jeweils eine Ähnlichkeit zwischen 3.6 und 4.2 gefunden, sofern die Submenge von COX2 bzw. P38 der Gesamt-Menge von P38 bzw. COX2, nicht aber THR, hinzugefügt ist. Hierbei ist jedoch zu beachten, dass auch zwischen den beiden Gesamtmengen („all_P38“ und „all_COX2“) eine Ähnlichkeit von 3.2 gefunden wird. Wie in Abschnitt 23.1.5 später noch analysiert und diskutiert, ist bei diesen Targets eine Kreuzreaktivität sehr wahrscheinlich. Somit werden ohne Berücksichtigung von Substruktur-Beziehungen relativ hohe Ähnlichkeiten gefunden, weil zusätzlich zu den jeweils künstlich hinzugefügten MCSs weitere MCSs in der Gesamtmenge beider Targets gleich sind.

Ebenfalls ist anhand der Abbildungen zu sehen, dass eine gewisse Varianz (v.a. in Abbildung 23.2) bei den Ähnlichkeitswerten der 10 verschiedenen Stichproben festzustellen ist. Diese ist auf die Gewichtung der unterschiedlichen MCS-Größen bei der Berechnung des gewichteten inSARa-TSim zurückzuführen.

Zusammenfassend lässt sich feststellen, dass es möglich ist insbesondere aufgrund der Berücksichtigung von Substruktur-Beziehungen auch Ähnlichkeiten (> 3.0) bei Vorhandensein von wenigen gleichen MCSs und einer großen Zahl unterschiedlicher MCSs in zwei großen MCS-Mengen zu erkennen. Dies ist sehr wichtig, da dies den Normalfall beim Vergleich der MCS-Mengen von zwei unterschiedlichen Targets darstellt.

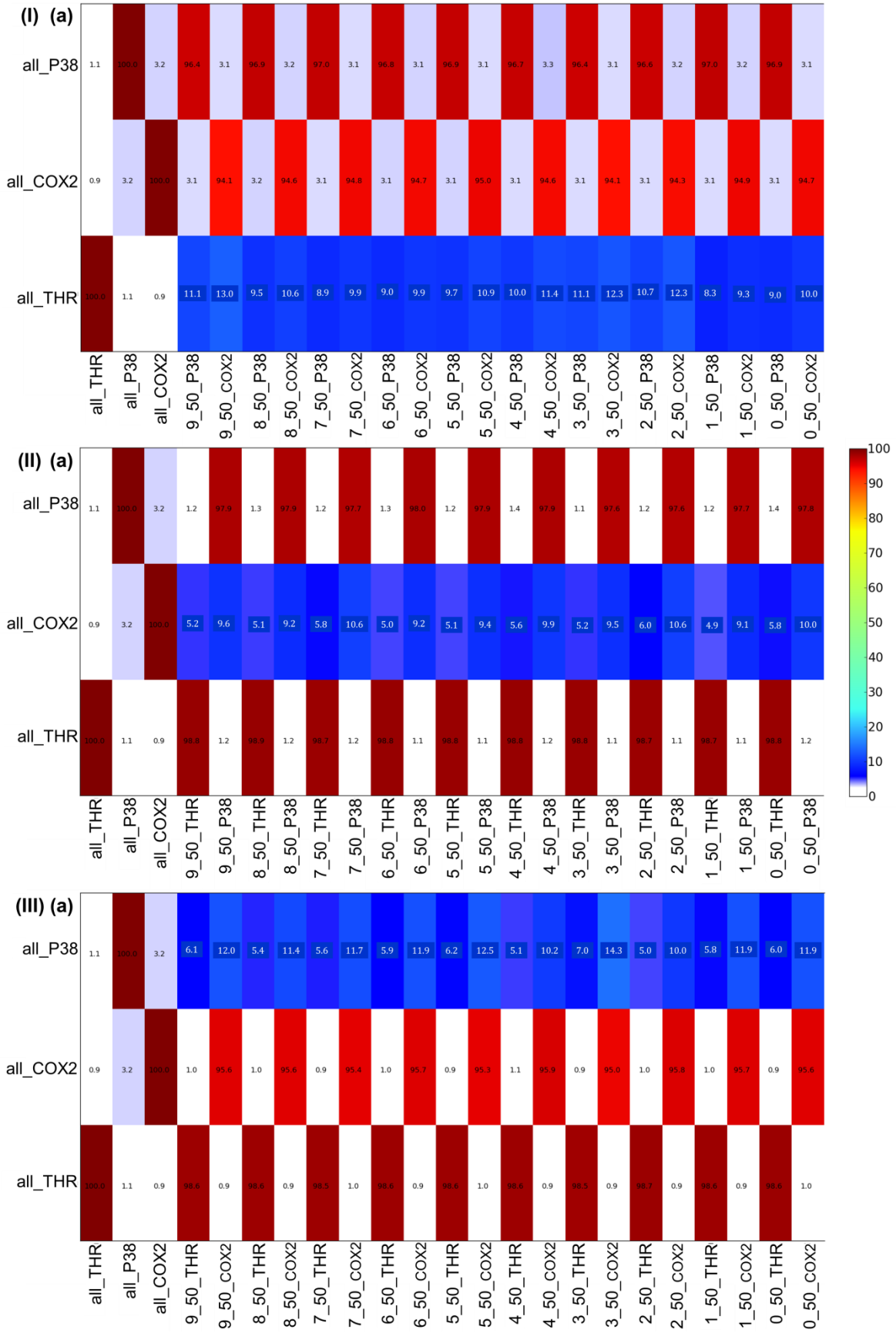


Abbildung 23.2. Ergebnisse des Wiederfindungstestes (gewichteter inSARa-TSim): (a) *mit* Berücksichtigung von Substruktur-Beziehungen: (I) *50 Thrombin-MCSs* in der COX2/P38-Gesamt-MCS-Menge, (II) *50 COX2-MCSs* in der THR/P38-Gesamt-MCS-Menge, (III) *50 P38-MCSs* in der THR/COX2-Gesamt-MCS-Menge. Details zu Abkürzungen siehe Text.

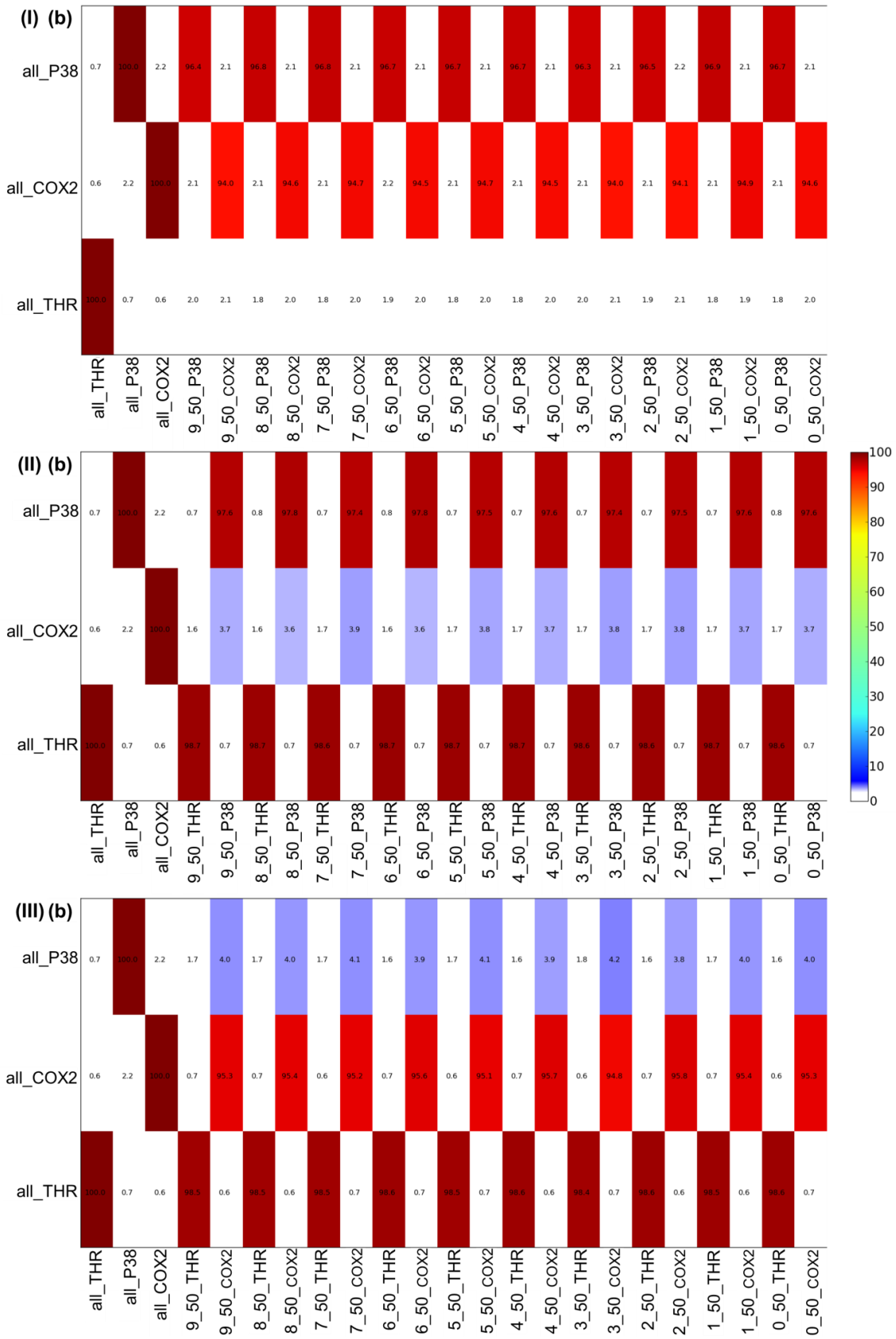


Abbildung 23.3. Ergebnisse des „Wiederfindungstestes“ (gewichteter inSARa-TSim): (b) *ohne* Berücksichtigung von Substruktur-Beziehungen: (I) *50 Thrombin-MCSs* in der COX2/P38-Gesamt-MCS-Menge, (II) *50 COX2-MCSs* in der THR/P38-Gesamt-MCS-Menge, (III) *50 P38-MCSs* in der THR/COX2-Gesamt-MCS-Menge. Details zu Abkürzungen siehe Text.

23.1.2. Ähnlichkeitskarten

In Abbildung 23.4 und Abbildung 23.5 sind die Ähnlichkeitskarten dargestellt, die aus dem Vergleich der 140 Targets unter Verwendung des ungewichteten bzw. gewichteten inSARa-TSim-Scores für die Ähnlichkeitsberechnung resultieren.

Durch Berücksichtigung der MCS-Größe bei der Ähnlichkeitsberechnung (vgl. Abbildung 23.5) ist eine deutliche Reduktion der gefärbten Flächen, d.h. Ähnlichkeit oberhalb des Schwellenwertes, in der Ähnlichkeitskarte (im Vergleich zu Abbildung 23.4) zu beobachten. Durch die Gewichtung ist somit eine Differenzierung von im Hinblick auf Polypharmakologie und die Arzneistoffentwicklung bedeutsameren (d.h. Ähnlichkeit beruht auf großen gemeinsame MCSs) von wenig bedeutsameren (d.h. Ähnlichkeit beruht auf einer großen Zahl kleiner gemeinsamer MCSs) Ähnlichkeitsbeziehungen möglich.

Es ist deutlich erkennbar, dass einige Targets zu einer Vielzahl von anderen Targets eine Ähnlichkeit aufweisen (z.B. Serotonin-, Dopamin- und Muskarin-Rezeptoren, sowie SERT, DAT und NET), wohingegen für einzelne Targets (z.B. B_2-Rezeptor, HMG-CoA-Reduktase, SLGT-2) keinerlei Ähnlichkeitsbeziehungen hergestellt werden können. Erwartungsgemäß weisen v.a. verschiedene Rezeptor-Subtypen (z.B. ET-α/β, mGlut-1/5, Orexin-1/2, CB-1/2, Estrogen-α/β) oder verschiedene Isoformen von Enzymen (z.B. COX-1/2, MAO-A/B, PDEs) hohe Ähnlichkeiten untereinander auf. Aufgrund der hohen Struktur-Homologie sind hier Kreuzreaktivitäten zu erwarten, sodass bei der Entwicklung neuer Arzneistoffe zur Bestimmung der Selektivität der Arzneistoff an den offensichtlich ähnlichen Targets i.d.R. ebenfalls getestet wird.

Zum detaillierten Studium einzelner Ähnlichkeitswerte sei auf Abbildung 26.5 bis Abbildung 26.7 im Anhang verwiesen, die eine Vergrößerung der Abbildung 23.5 darstellen.

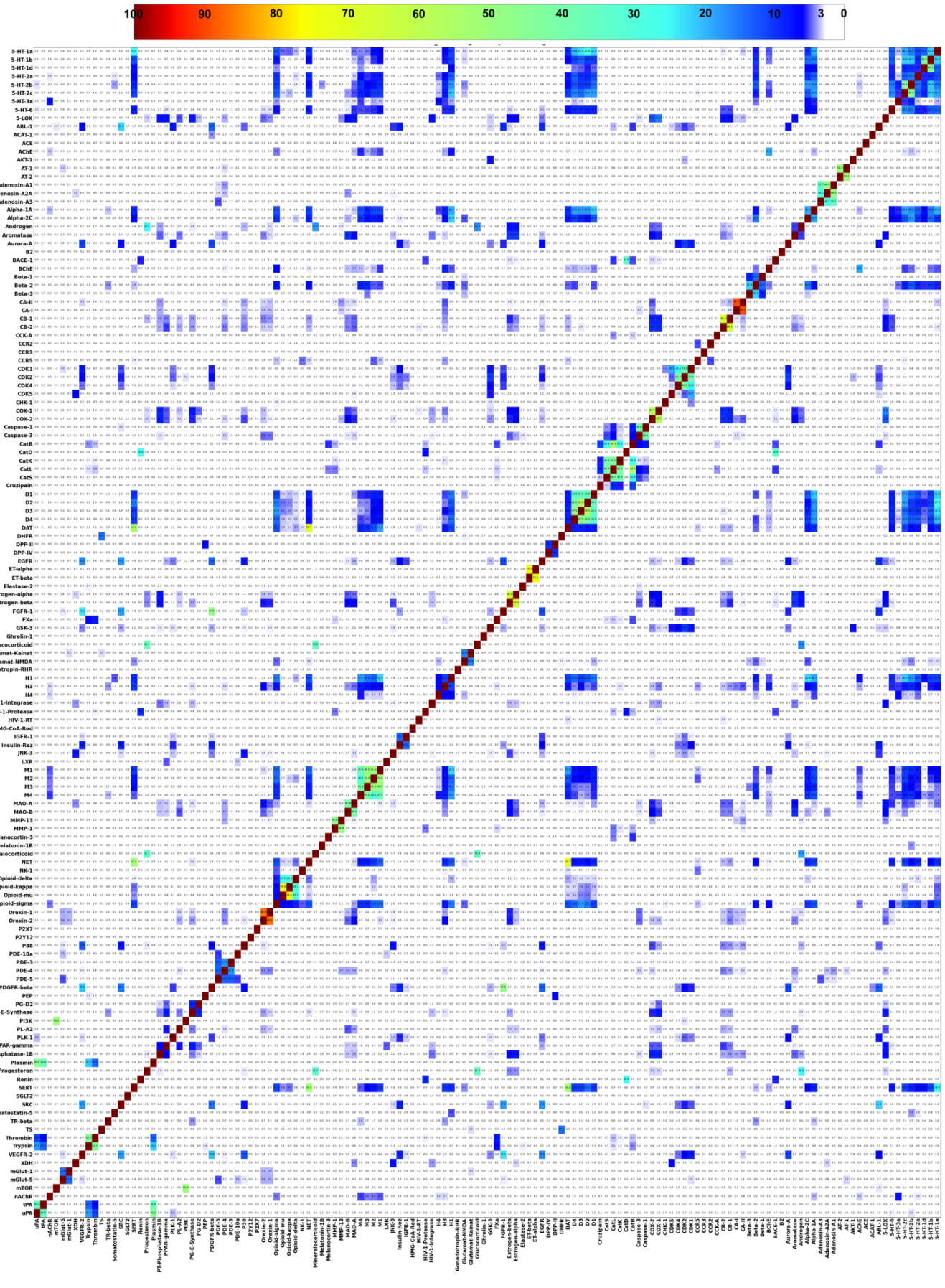


Abbildung 23.4. Ähnlichkeitskarte der inSARa-Netzwerke der 140 Targets (ungewichteter inSARa-TSim). weiß = geringe paarweise Ähnlichkeit, Ähnlichkeits-Zunahme von blau nach rot (vgl. Legende).

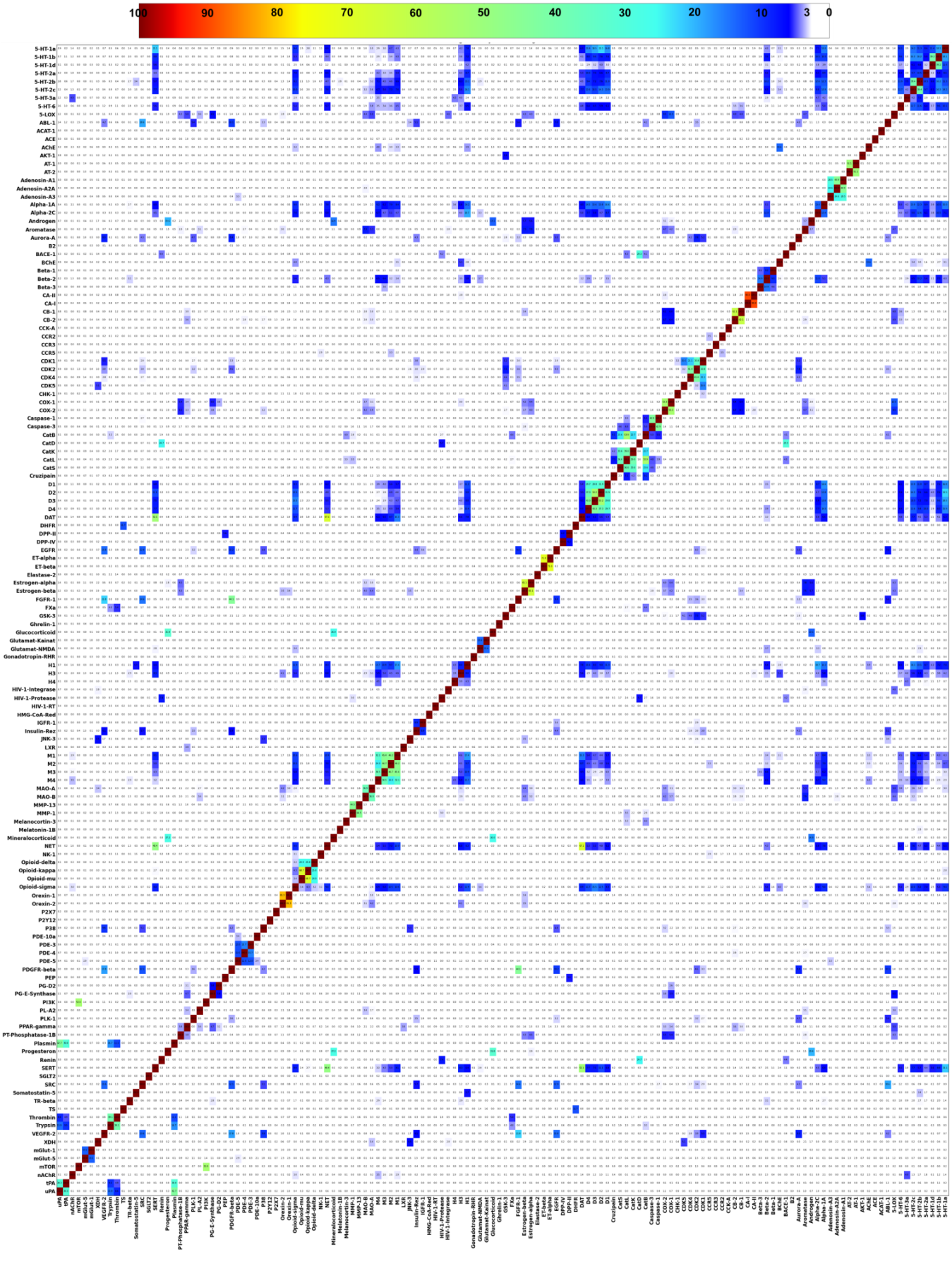


Abbildung 23.5. Ähnlichkeitskarte der inSARa-Netzwerke der 140 Targets (gewichteter inSARa-TSim). weiß = geringe paarweise Ähnlichkeit, Ähnlichkeits-Zunahme von blau nach rot (vgl. Legende).

23.1.3. Analyse von Target-Ähnlichkeiten mittels Schwellenwert-Netzwerk

In Abbildung 23.6 ist das aus Abbildung 23.5 für einen Schwellenwert von 3.0 resultierende Schwellenwert-Netzwerk gezeigt.

Analyse der Clusterbildung auf Basis von Target-Klassen-Information

In Abbildung 23.6 ist zu erkennen, dass zumeist Knoten gleicher Farbe durch Kanten verbunden sind, d.h. Targets gleicher Protein-Familien zusammengruppiert sind. Hier zeichnen sich relativ deutlich Cluster ab. So findet man beispielsweise unten in Abbildung 23.20 einen großen Protease-Cluster (hellblaue Knoten), in der Mitte links einen großen Kinase-Cluster (rote Knoten), in der Mitte einen Cluster von nukleären Rezeptoren (gelbe Knoten) und einigen Enzymen (dunkelblaue Knoten), sowie rechts oben einen großen GPCR-Cluster (grüne Knoten). Daneben sind noch zahlreiche Einzelknoten (v.a. grüne GPCR-Knoten und dunkelblaue Enzym-Knoten) bzw. kleinere Einzelcluster (z.B. die Kinasen mTor und PI3K, die PDEs und Adenosin-Rezeptoren, die Thymidilat-Synthase und die DHFR oder aber die Serin-Proteasen PEP und DPP-II/IV) zu erkennen. Die Proteasen weisen zusammen mit den Kinasen die wenigsten Proteinfamilien-übergreifenden Verknüpfungen auf. Die GPCRs und anderen Enzyme weisen hingegen vielfältige Proteinfamilien-übergreifende Verknüpfungen auf. Die Transporter (SERT, NET, DAT) und ligandengesteuerten Ionenkanäle (5-HT_{3a}, nAChR, Glutamat-NMDA) weisen oftmals Ähnlichkeiten zu den GPCRs (aminerg) auf.

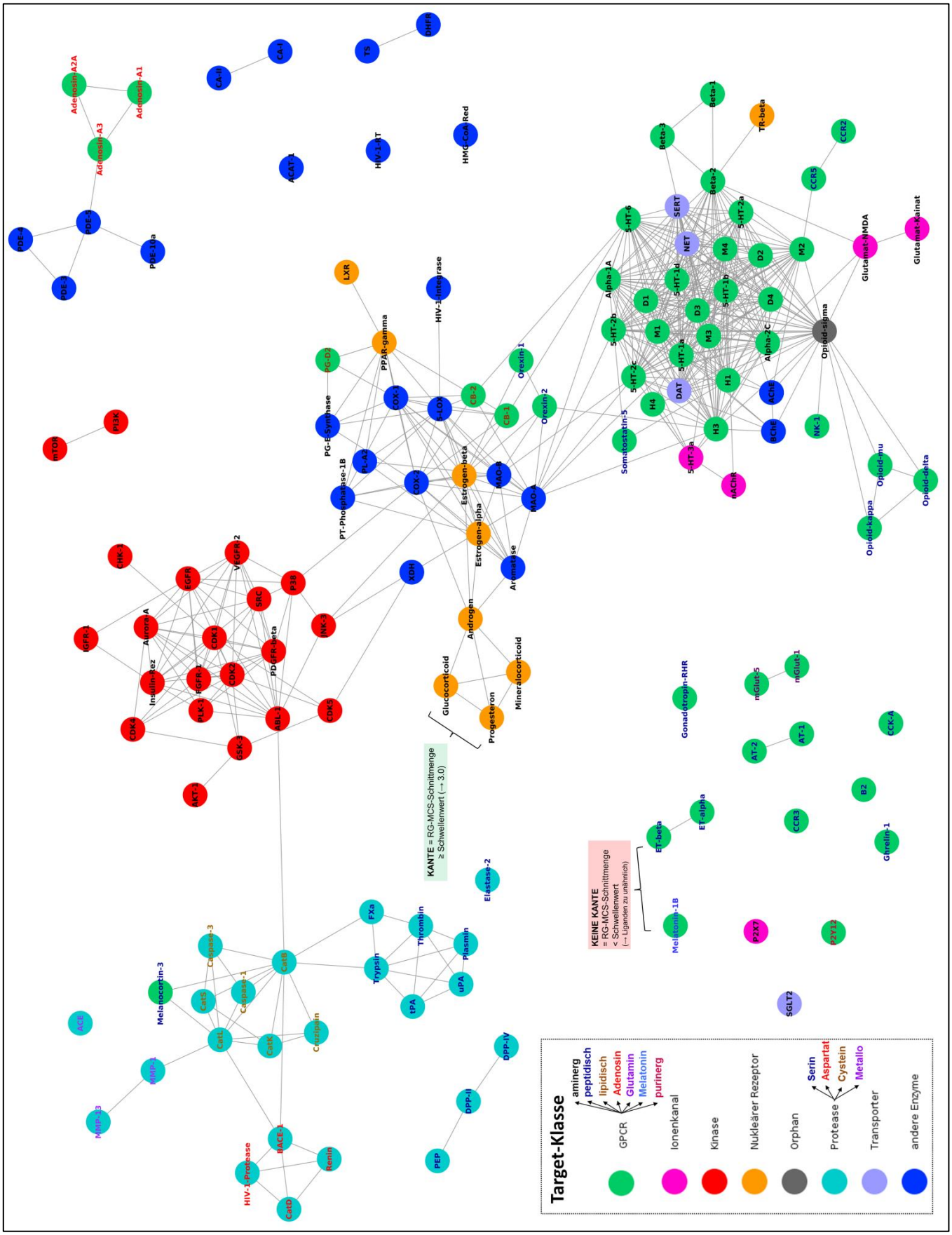


Abbildung 23.6. Schwellenwert-Netzwerk, das aus dem Ähnlichkeitsvergleich der inSARa-Netzwerke der 140 BindingDB-Targets resultiert (Schwellenwert = 3.0, gewichteter inSARa-TSim). Knoten-Farbe gemäß Target-Klasse, Label-Farbe gemäß Subklasse, vgl. Legende).

Beurteilung der inSARa-basierten Target-Beziehungen auf Basis von Subklassen-Informationen

a) GPCRs

GPCRs können anhand der endogenen Liganden in verschiedene Subklassen (biogene Amine, d.h. Histamin, Acetylcholin, Serotonin, Noradrenalin, Dopamin, = aminerg, Peptide = peptidisch, Fettsäure-Derivate = lipidisch, ADP/ATP = purinerg, Adenosin, Melatonin, Glutamin) eingeteilt werden.[349, 415] Berücksichtigt man zusätzlich diese Subklassen (in Abbildung 23.20 durch die Farbe des Knoten-Labels gekennzeichnet, vgl. Legende), ist festzustellen, dass meist Targets gleicher Subklassen miteinander verbunden sind bzw. Cluster bilden. Während die aminergen GPCRs große Ähnlichkeit untereinander aufweisen (vgl. zahlreiche Verknüpfungen im GPCR-Cluster), stellen die peptidischen GPCRs meist Einzelknoten dar bzw. tauchen verstreut im Netzwerk auf. Die lipidischen GPCRs (Cannabinoid-Rezeptor und Prostaglandin-D2-Rezeptor) finden sich in dem Cluster wieder, wo zahlreiche Targets des Steroid- oder Fettsäure-Stoffwechsel zu finden sind. Die anderen Subklassen bilden jeweils einzelne Cluster.

Die hohe Verknüpfungsdichte innerhalb der aminergen GPCRs ist aufgrund bekannter privilegierter Strukturen zu erwarten. Auch spiegelt sich dieses Bild im biologischen Aktivitätsprofil einer Vielzahl von v.a. antipsychotischen oder antidepressiven Arzneistoffen wieder (z.B. Musterbeispiel das atypische Neuroleptikum Clozapin).[463] Ein typisches für Promiskuität an aminergen GPCRs (aber auch z.B. für Monoamin-Transporter) bekanntes Motiv, dass sich auch in den gemeinsamen RG-MCSs wiederfindet, ist: Basisches, protoniertes Amin (PI-Eigenschaft = Nb) über Linker bestimmter Länge (Zn) mit meist zwei oder mehr Aromaten (z.B. Sc) verknüpft.[307, 464]

b) Proteasen

Die analysierten Proteasen lassen sich ebenfalls basierend auf bestimmten Aminosäuren, die am Katalysemechanismus beteiligt sind, in 4 Haupt-Subklassen (Serin-, Aspartat, Cystein- und Metalloproteasen) einteilen.[465] Auch hier lässt sich feststellen, dass diese Subklassen jeweils kleine eigene Sub-Cluster im Netzwerk bilden (vgl. Farbe des Knoten-Labels in Abbildung 23.20). Bei den Serinproteasen lassen sich grob zwei Cluster (FXa, Thrombin etc. vs. PEP/DPPs) unterscheiden, bei den Metalloproteasen ebenfalls (ACE vs. die MMPs).

c) Kinasen

Die Kinasen bilden einen großen Cluster, wobei die Targets i.d.R. vielfältige Verknüpfungen untereinander aufweisen. Im Vergleich zum Cluster der aminergen GPCRs ist die Verknüpfungsdichte im Kinasen-Cluster etwas geringer. AKT-1 und CHK-1 bilden hierbei eine Ausnahme und sind nur mit einer weiteren Kinase verknüpft. Die beiden Kinasen mTor und PI3K bilden einen separaten Cluster.

Da wie für die GPCRs auch für die Kinasen ebenfalls eine Reihe privilegierter Strukturmotive beschrieben sind[307, 446], ist die hohe Ähnlichkeit untereinander zu erwarten. So gibt es z.B. das „Hinge-Bindungs-Motiv“ (vgl. Kapitel 20.2.1), das strukturelle Merkmale, die für die Bindung an die in allen Kinasen hochkonservierte Scharnier-Region beschreibt. Bekanntermaßen ist die Entwicklung von Kinase-Inhibitoren, die kompetitiv in der ATP-Bindetasche binden, in Bezug auf Selektivität (v.a. bei Typ I Kinase-Inhibitoren)[441] gegenüber anderen Kinasen aufgrund der hohen Sequenzhomologie in dieser Region sehr schwierig.[466] Eine Analyse der gemeinsamen MCSs mit unerwünschten Targets könnte bei der Entwicklung selektiverer Moleküle bzw. beim Erkennen (un)selektiver molekularer Merkmale helfen.

d) Weitere Enzyme

Bei den übrigen Enzymen ergibt sich kein so einheitliches Bild wie bei den Proteasen, Kinasen oder den meisten anderen Target-Klassen. In Anbetracht der unterschiedlichen katalytischen Funktion (vgl. Klassifizierung nach EC-Code in Tabelle 17.2 nach IUBMB Nomenklatur[419]) und unterschiedlichen Substrate ist diese Aufspaltung in verschiedene kleine Cluster leicht verständlich. Verschiedene Isoformen, die i.d.R. ebenfalls hohe Sequenz-Ähnlichkeit aufweisen, sind erwartungsgemäß aufgrund ähnlicher Liganden im Netzwerk miteinander verknüpft (z.B. PDEs, CA-I/II, COX-1/2, MAO-A/B). Die AChE und BChE, die beide die Hydrolyse von Cholin-Estern katalysieren sind ebenfalls verknüpft. Die Thymidilat-Synthase und die Dihydrofolat-Reduktase weisen ebenfalls hohe Ähnlichkeit und bilden einen eigenen Cluster, obwohl sie nach EC-System aufgrund ihrer katalytischen Funktionen in sehr verschiedene Klassen eingeteilt werden. Eine Erklärung für die Ähnlichkeit ist, dass die Liganden beider Targets Folsäure-Analoga darstellen.[345] Als Beispiel-Arzneistoff ist Pemetrexed zu nennen, der durch die duale Inhibition beider Targets an mehreren Stellen der DNA-Synthese eingreift und so eine effektive Therapie bestimmter Tumorformen ermöglicht.[467] Ansonsten lässt sich feststellen, dass die Enzyme vielfältige Verknüpfungen zu Targets anderer Klassen aufweisen können, die zumeist durch ähnliche Eigenschaften der Liganden bedingt sind (vgl. auch Abschnitt 23.1.4). Der in Abbildung 23.8 braun umrandete Cluster, in dem auch zahlreiche Enzyme zu finden sind, lässt sich dadurch erklären, dass diese Targets in den Lipid-Stoffwechsel (Steroide, Fettsäure-Derivate, Prostaglandine etc.) involviert sind. Zahlreiche Analysen zeigen ebenfalls, dass Moleküle hoher Lipophilie eine Neigung zu Promiskuität aufweisen.[307, 468]

e) Nukleäre Rezeptoren

Bei den nukleären Rezeptoren zeichnen sich drei Cluster ab. Die Steroidhormon-Rezeptoren, wobei die Estrogen-Rezeptoren über den Androgen-Rezeptor mit dem 3-Ketosteroid-Cluster (bestehend aus dem Glucocorticoid-, Mineralocorticoid-, Progesteron-, Androgen-Rezeptor) verknüpft ist. Der PPAR-γ- und der LXR-Rezeptor bilden eine separate Gruppe. Der ebenfalls zu den RXR-Heterodimeren gehörende Thyroid-Hormon-Rezeptor-β bildet einen von diesen Gruppierungen weit entfernten separaten Cluster. Aufgrund der endogenen Liganden wäre eine Ähnlichkeit des LXR-Rezeptors (Oxysterole) zum Steroidhormon-Rezeptor eher zu erwarten gewesen als zu PPAR-γ (Fettsäuren und Eicosanoide).

f) Transporter

Die Transporter SERT, DAT und NET weisen eine hohe Ähnlichkeit untereinander auf. Zum SLGT2-Transporter ist hingegen keine Ähnlichkeit festzustellen. Während SERT, NET und DAT für den Transport der biogenen Amine Serotonin, Dopamin und Noradrenalin verantwortlich sind, ist der Natrium-abhängige Glucose-Transporter für die Glucose-Resorption im proximalen Tubulus verantwortlich. Durch die (Un-)Ähnlichkeit der endogenen Liganden ist die Ähnlichkeit zwischen den Monoamin-Transportern untereinander bzw. Unähnlichkeit zum SLGT2-Transporter leicht erklärbar. Die gefundene Ähnlichkeit ist auch von klinischer Relevanz: Für zahlreiche Arzneistoffe aus dem Bereich der Antidepressiva/Neuroleptika[463] oder Psychostimulantien (z.B. Cocain)[469] ist die Wirkung auf gleichzeitige Inhibition dieser Monoamin-Transporter-Proteine zurückzuführen.

g) Liganden-gesteuerte Ionenkanäle

Die Liganden-gesteuerten Ionenkanäle weisen untereinander nur partiell Ähnlichkeiten auf (5-HT_3/nAChR, Glutamat-NMDA/Kainat). Der Purinorezeptor P2X7 bildet einen Einzelknoten im Netzwerk. Auf Basis der endogenen Liganden ist diese Gruppierung sinnvoll. Die ionotropen Glutamat-Rezeptoren weisen eine hohe Affinität für die Aminosäure Glutamat auf, während die biogenen Amine Serotonin und Acetylcholin endogene Liganden am ionotropen Serotonin-Rezeptor bzw. nikotinischen Acetylcholin-Rezeptor darstellen. Die pharmakologische Relevanz wird auch durch experimentelle Testungen verschiedener 5-HT_3-Antagonisten bestätigt (z.B. Tropisetron: K_i(5-HT_3) = 5.3 nM, K_i = 6.9 nM (nAChR)).[470] ATP, der endogene Ligand am ionotropen Purinorezeptor, unterscheidet sich strukturell deutlich von den vorgenannten Liganden.

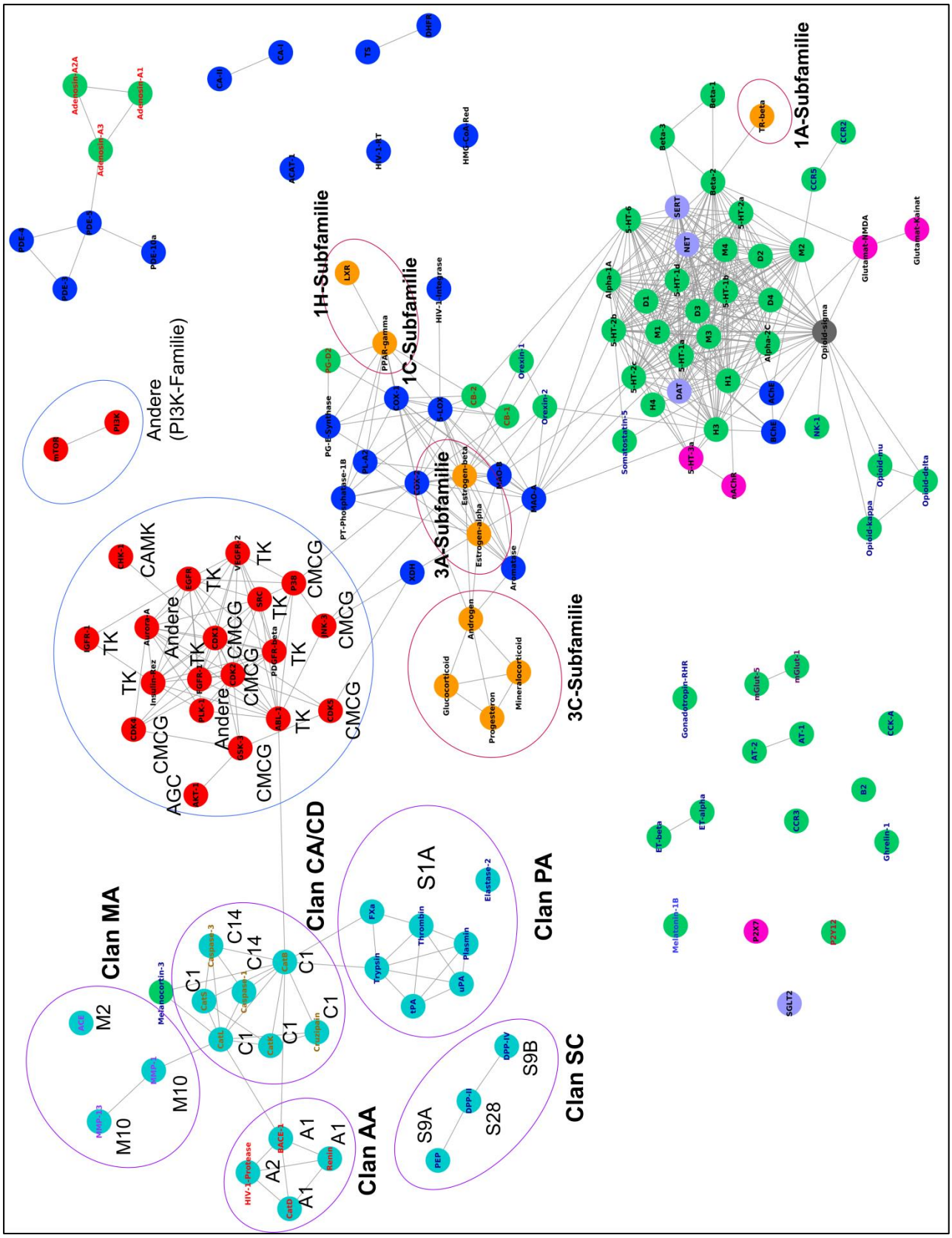


Abbildung 23.7. Schwellenwert-Netzwerk aus Abbildung 23.6 im Vergleich mit Sequenz-basierter Ähnlichkeit (Annotation analog zu Tabelle 17.2, Kinasen = blau markiert, Proteasen = lila markiert, nukleäre Rezeptoren = rot markiert). Für Details siehe Text.

Vergleich mit Sequenz-basierter Ähnlichkeit

a) GPCRs

Obwohl im Allgemeinen zumeist eine höhere Ähnlichkeit auch auf Basis des inSARa-Netzwerk-Vergleichs zwischen Target mit hoher Sequenz-Ähnlichkeit gefunden wird, so lassen sich (wie auch in Analysen anderer Gruppen festgestellt[349, 343, 384]) ebenfalls deutliche Abweichungen feststellen. Als Beispiel sind hier der Somatostatin-5-Rezeptor und H_1-Rezeptor zu nennen, die nur eine geringe Bindetaschen-Sequenz-Ähnlichkeit (33%) aufweisen[349]. Auf Basis verschiedener Sequenz-Homologie-Analysen[343, 349] wäre eine erhöhte Ähnlichkeit zu den Opioid-Rezeptoren zu erwarten, die in der ligandbasierten Analyse nicht bestätigt werden kann. Ein weiteres Beispiel stellt der CCR5-Rezeptor dar, der zu dem M_2-Rezeptor fast eine genauso große inSARa-Ähnlichkeit aufweist wie zum CCR2-Rezeptor. Während die Sequenzhomologie zwischen dem CCR2-Rezeptor und CCR5-Rezeptor sehr hoch ist, ist die Bindetaschen-Sequenz-Ähnlichkeit zum M_2-Rezeptor nur sehr gering (16%)[349]. Aufgrund von Sequenzhomologie wäre statt zum M_2-Rezeptor zusätzlich eine Ähnlichkeit zum CCR3-Rezeptor zu erwarten gewesen.[343, 349] Für die beiden Beispiele zeigt die SEA-basierte GPCR-Analyse von LIN et al.[349] ähnliche Target-Beziehungen wie der inSARa-basierte Ansatz, während die Ergebnisse der Substruktur-basierte Analyse von VAN DER HORST et al.[343] deutlich differieren.

b) Proteasen

Die MEROPS-Datenbank liefert eine hierarchische, Sequenz-basierte Klassifikation von Proteasen auf Grundlage der Aminosäuren der Peptidase-Domain.[338] In Tabelle 17.2 und Abbildung 23.7 (lila Markierung) sind die entsprechenden (Sub-)Familien-/Clan-Annotationen für die analysierten Proteasen zu finden. Der Buchstabe gibt dabei immer den katalytischen Typ an, gefolgt von der Familien- und ggf. Subfamilien-Einteilung (z.B. FXa = S1A = Serinprotease, Familie 1, Subfamilie 1A).[417] Familien bilden Proteasen, die eine statistisch signifikante Ähnlichkeit in der Aminosäuresequenz aufweisen, bei Subfamilien ist diese Ähnlichkeit noch höher.[417] Homologe Familien lassen sich abermals zu Clans zusammenfassen.[417]

Vergleicht man die inSARa-basierten Beziehungen mit der Sequenz-Ähnlichkeit der Protease-Domain, so lassen sich deutliche Analogien feststellen. Ebenfalls liefert die Sequenz-Ähnlichkeit eine Begründung für die Subgruppenbildung (FXa und andere/PEP und DPPs) bei den Serinproteasen (S1- und S9/28-Familie bzw. Clan PA und SC). Die ebenfalls zur S1A-Familie gehörende Elastase-2 (Einzelknoten im Schwellenwert-Netzwerk) weist die auf Basis der inSARa-TSim-Werte die größte Ähnlichkeit zu der S1A-Protease Thrombin (2.1) auf. Ansonsten sind nur deutlich geringe Ähnlichkeiten zu einigen Cysteinproteasen festzustellen. Auch lassen sich die zwei Cluster innerhalb der Metalloproteasen (ACE/MMPs) auf unterschiedliche Familien (M2 und M10) zurückführen. Innerhalb der anderen Protease-Klassen lassen diese Aufspaltungen analog der Sequenzähnlichkeiten nicht beobachten.

c) Kinasen

Auf Basis von Sequenz-Ähnlichkeit der Protein-Kinase-Domain (vgl. MANNING[336]) lassen sich die Kinasen in verschiedene Gruppen/Familien einordnen. In Tabelle 17.2 und Abbildung 23.7 (blaue Markierung) sind diese Gruppen-Annotationen für die Target-Klasse der Kinasen zu finden. Es lässt sich hierbei feststellen, dass im Allgemeinen auch im inSARa-basierten Schwellenwert-Netzwerk Targets einer Gruppe miteinander verknüpft sind. Jedoch sind auch oft Gruppen-übergreifende Verknüpfungen zu finden. Die beiden Kinasen AKT-1 (AGC-Gruppe) und CHK-1 (CAMK-Gruppe), die sich nicht wie die Mehrheit der Kinasen der TK- oder CMCG-Gruppe zuordnen lassen, weisen jedoch im Gegensatz zu den anderen Kinasen nur eine Verknüpfung im Netzwerk auf. Die Abspaltung des mTor/PI3K-Clusters lässt sich aufgrund von Sequenz-Ähnlichkeit ebenfalls begründen. Beide Targets weisen keine große Sequenz-Ähnlichkeit zu den großen Kinase-Gruppen auf. Sie bilden zusammen die PI3K-Familie.

d) Nukleäre Rezeptoren

Bei den nukleären Rezeptoren spiegeln die inSARa-Netzwerk-basierten Ähnlichkeits-Beziehungen die Sequenz-Ähnlichkeit innerhalb der Familie wider (vgl. auch Subfamilien-Annotationen in Tabelle 17.2 und Abbildung 23.7 (rote Markierung)).[418] So zeigen der PPAR-γ-Rezeptor und der LXR-Rezeptor untereinander eine höhere Sequenz-Ähnlichkeit (1C- und 1H-Subfamilie) als die Steroidhormon-Rezeptoren (3-Subfamilie).[471] Die Untergruppenbildung innerhalb der Steroidhormon-Rezeptor-Familie wird auch durch die phylogenetische Ähnlichkeit bestätigt (3A- und 3C-Subfamilie).[471]

Zusammenfassung: Vergleich mit Sequenz-basierter Ähnlichkeit

Zusammenfassend lässt sich wie auch bei Analysen anderer Gruppen[343, 345, 349, 384] feststellen, dass die Ergebnisse der ligandbasierten Target-Analyse im Allgemeinen auch Sequenz-Ähnlichkeiten widerspiegeln. Proteine, die zum gleichen Clan/(Sub-) Familie/Gruppe gehören, haben zumeist ähnliche Bindetaschen und weisen in der Regel somit auch Gemeinsamkeiten in Bezug auf die Liganden (pharmakophore Eigenschaften) auf. Es lassen sich jedoch im ligandbasierten Target-Netzwerk in einer Vielzahl von Fällen auch Unterschiede zur Sequenz-basierten Analyse erkennen, v.a. wenn Targets verschiedener Target-Klassen verknüpft sind. Während mit dieser ligandbasierten Analyse auf einfache Weise auch Targets unterschiedlicher Target-Klassen verglichen werden können, entstehen bei Sequenz-Analysen deutliche Schwierigkeiten (vgl. Kapitel 8.2). Ligandbasierte Ähnlichkeit ist zudem im Gegensatz zu Sequenz-basierter häufig von klinischer Relevanz, da bei hoher Liganden-Ähnlichkeit von potentieller Bindung der Liganden an das jeweils andere Targets ausgegangen werden kann. Diese Target-Beziehungen sind sehr wertvoll für die Erklärung pharmakologischer (un-)erwünschter Wirkungen. Ligandbasierte Analysen stellen somit eine wichtige Ergänzung zu Sequenz-basierten Verfahren dar.

Vergleich mit ligandbasierten Target-Netzwerken anderer Gruppen

Aufgrund der unterschiedlichen Datengrundlage sind direkte Vergleiche mit Analysen anderer Gruppen schwierig. Jedoch lässt sich grob vergleichen, ob sich ähnliche Trends in den Ergebnissen bzw. Target-Netzwerken abzeichnen.

In den Target-Netzwerken von PAOLINI et al.[342] und KEISER et al.[345] (vgl. Kapitel 8.2) zeigt sich ebenfalls wie im inSARa-basierten Netzwerk, dass bestimmte Target-Klassen einzelne Cluster bilden, die sich mehr oder weniger stark von den anderen Targets abgrenzen. Die Topologie der Netzwerke zeigt erwartungsgemäß jedoch aufgrund der unterschiedlichen Repräsentation und dem Ähnlichkeitsvergleich deutliche Differenzen. So bilden bei PAOLINI et al.[342] beispielsweise die einzelnen Subklassen der Proteasen ebenfalls eigene Cluster. Diese bilden aber nicht wie in Abbildung 23.6 einen gemeinsamen Protease-Cluster, sondern sind über andere Target-Klassen verknüpft. Die Analyse von PAOLINI et al.[342] findet eine hohe Intra- und Intertargetklassen-Promiskuität für aminerge GPCRs, die sich auch durch die inSARa-basierte Analyse bestätigen lässt. Viele weitere Klassen (z.B. Kinasen, einige Protease-Klassen und die peptidischen GPCRs) weisen zudem eine hohe Intratargetklassen-Promiskuität auf.[342] Mit Ausnahme der peptidischen GPCRs, die in Abbildung 23.6 meist keine Verknüpfung zu anderen peptidischen GPCRs oder Targets aufweisen, lassen sich diese Ergebnisse durch die inSARa-basierte Analyse ebenfalls bestätigen. Eine mögliche Ursache für das fehlende Erkennen von Ähnlichkeitsbeziehungen könnte u.U. eine für peptidische Liganden nicht-optimale RG-Definition sein (vgl. auch Abschnitt 23.1.5). Vergleicht man den Substruktur-basierten Ansatz von SUTHERLAND et al.[344] mit dem inSARa-basierten Ansatz oder der Analyse von PAOLINI et al.[342] so lassen sich deutliche Unterschiede feststellen. So finden SUTHERLAND et al.[344] für aminerge GPCRs nur eine geringe durchschnittliche Fragment-Ähnlichkeit. Stattdessen wird innerhalb einer Protease-Klasse und zu anderen Protease-Klassen, sowie zu peptidischen GPCRs eine hohe durchschnittliche Fragment-Ähnlichkeit gefunden.[344] Ebenfalls wird für viele Target-Klassen (im Gegensatz zu PAOLINI et al.[342], KEISER et al.[345] und dem inSARa-basierten Ansatz) keine hohe Intratargetklassen-Ähnlichkeit (z.B. für Kinasen) gefunden.[344] Eine Erklärung für das fehlende Erkennen der Ähnlichkeit im Vergleich zum RG-MCS-basierten Ansatz könnte die fehlende Abstraktion auf pharmakophore Eigenschaften sein. Auf Basis des exakten Vergleiches der Molekülstruktur (z.B. variierende Linker, unterschiedliche PI-Gruppen) können die vorhandenen Ähnlichkeiten scheinbar nur schwer erkannt werden. Die hohe Ähnlichkeit zwischen peptischen GPCRs und Proteasen ist eher „artifizieller Natur“ und pharmakologisch wenig relevant, da bei dem Fragment-basierten Ansatz die unspezifischen peptidischen Fragmente, die in Molekülen beider Gruppen häufig vorkommen, eine hohe Ähnlichkeit verursachen.

Zusammenfassend lässt sich feststellen, dass Ähnlichkeiten, aber auch Unterschiede beim Vergleich der verschiedenen Analysen identifiziert werden können. Neben der Abhängigkeit von zugrunde liegenden Daten sind auch methodische Unterschiede als Ursachen für die Unterschiede zu berücksichtigen. Da jede Methode Stärken (Substruktur-basiert z.B. Interpretierbarkeit im Vergleich zur FP-basierten Analyse) und Schwächen (Substruktur-basiert z.B. mangelhaftes Erkennen von Gemeinsamkeiten) aufweist, komplementieren sich die Verfahren. Die Kombination der Ergebnisse verschiedener Analysen kann somit sehr hilfreich dafür sein, ein vollständigeres Bild aller pharmakologischen Beziehungen zu erhalten. Im Vergleich mit SUTHERLAND et al.[344] erscheint die Verwendung von RGs ein vielsprechender Ansatz für verbesserte, intuitive Substruktur-basierte Analysen zu sein.

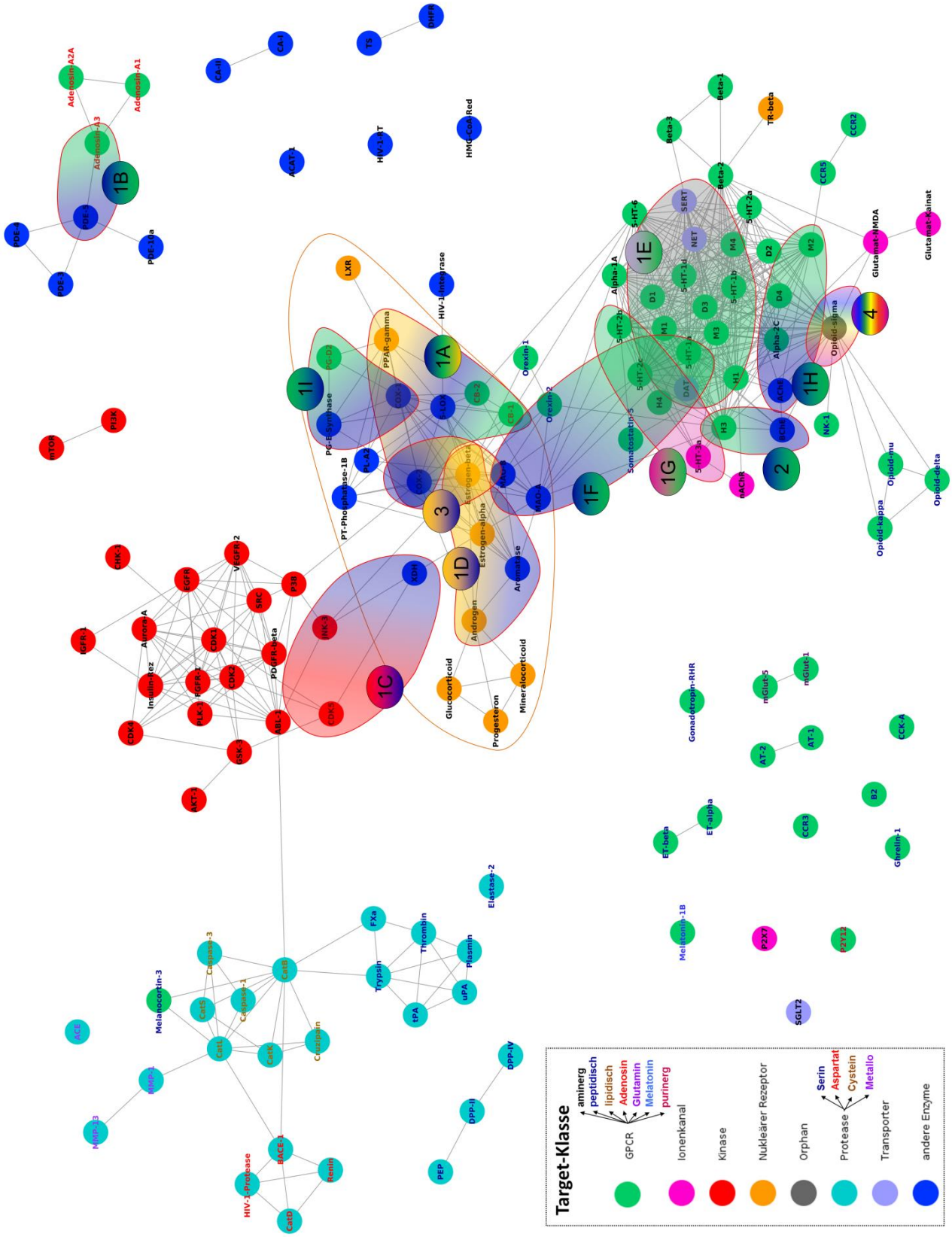


Abbildung 23.8. Schwellenwert-Netzwerk aus Abbildung 23.6 mit ausgewählten Kreuzreaktivitäten (Farbe der Markierung gemäß Target-Klassen der involvierten Targets). Für Details bezüglich der nummerierten Beispiele vgl. Text in Abschnitt 23.1.4)

23.1.4. Validierung potentieller Kreuzreaktivitäten (Literaturrecherche)

In dem Netzwerk in Abbildung 23.6 sind nicht nur Targets gleicher Protein-Klassen miteinander verknüpft, sondern es sind auch eine Reihe von unerwarteten Verknüpfungen zu beobachten, sogenannte potentielle „Kreuzreaktivitäten“. Im Folgenden werden einige dieser Beispiele ausgewählt (in Abbildung 23.8 markiert und nummeriert) und Ursachen für diese Kreuzreaktivitäten aufgezeigt und diskutiert.

1.) Ähnliche endogene Liganden

A) „Off-Target-Hotspot“ (COX, CB und PPAR-γ)

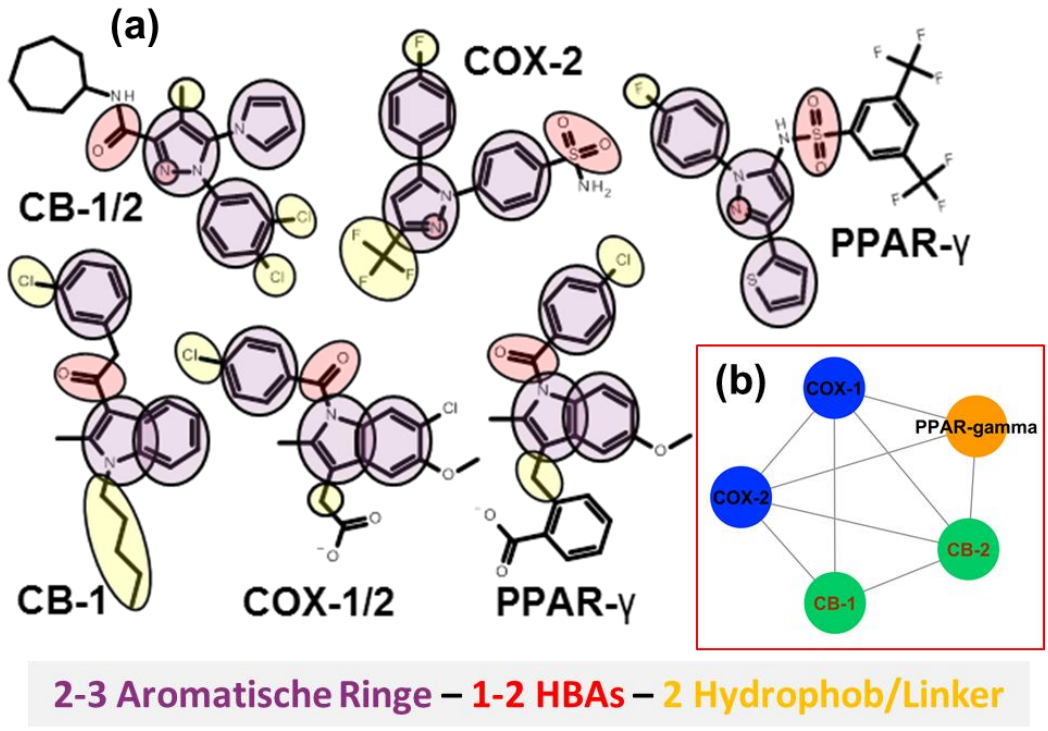


Abbildung 23.9. Beispiel für einen Off-Target-Hotspot aus Cyclooxygenase (COX), Cannabinoid-Rezeptor (CB) und Peroxisom-Proliferator-aktivierter Rezeptor gamma (PPAR-γ): Enzym – GPCR – nukleärer Rezeptor. (a) Liganden mit ähnlichem Grundgerüst, farbig markiert typische gemeinsame pharmakophore Eigenschaften. (b) Ausschnitt aus Schwellenwert-Netzwerk.

Viele der unerwarteten Beziehungen lassen sich aufgrund der endogenen Liganden erklären. So auch der in Abbildung 23.9 (b) gezeigte „Off-Target-Hotspot“. Hier sind Targets von drei verschiedenen Protein-Klassen miteinander verknüpft (GPCR, Enzym und nuklärer Rezeptor). In Abbildung 23.9 (a) sind Liganden der entsprechenden Targets gezeigt. Die Liganden weisen jeweils sehr ähnliche molekulare Eigenschaften auf. Das typische pharmakophore Motiv, das sich aus der gemeinsamen MCS-Menge der drei Targets ableiten lässt sind 2-3 aromatische Ringe, 1-2 H-Brücken-Akzeptoren und 2 hydrophobe Eigenschaften. Schaut man sich die endogenen Liganden der Targets an (COX: u.a. Arachidonsäure, CB: u.a. Anandamid, PPAR-γ: Eicosanoide, Fettsäuren und

Prostaglandine), so handelt es sich hierbei um langkettige Fettsäure-Derivate. Das gemeinsame pharmakophore Motiv stellt somit eine Mimikry dieser endogenen Liganden dar.

Zur Validierung dieser Kreuzreaktivität wurde nach in der Literatur beschriebenen experimenteller Bestätigung von biologischer Aktivität eines Liganden an dem potentiellen Off-Target gesucht. Die Ergebnisse dieser Recherche und die sich ergebenden pharmakologischen bzw. therapeutischen Konsequenzen sind im Folgenden zusammengefasst:

a) COX und CB-Rezeptor: Kreuzreaktivität anzunehmen

Es sind widersprüchliche Meinungen in der Literatur zu finden, jedoch kann eine Interaktion in einigen Fällen angenommen werden.[472] Der COX-Inhibitor Pravadoline (strukturell ähnlich zu Indometacin) ist im nanomolaren Konzentrationsbereich ein CB-Agonist.[473] Es gibt Hinweise, dass die analgetische Wirkung des COX-2 Inhibitors Celecoxib über das Cannabinoid-System vermittelt wird.[474] Die Wirkung des COX-2-Inhibitors Nimesulid ist zum Teil über den CB-Rezeptor vermittelt.[475] Es wären weitere experimentelle Testungen zur Klärung dieser Interaktion sinnvoll, da nichtsteroidale Antiphlogistika (NSARs) zu den am häufigsten eingesetzten Arzneimitteln gehören. Auch im Hinblick auf häufigen Missbrauch von NSARs im Leistungssport[476] sollten eine potentiell leistungssteigernde CB-Wirkung (Doping) untersucht werden. Für eine dualen CB-Agonismus/COX-2-Inhibition sind positive Effekte z.B. in der Therapie des Kolon-Karzinoms zu erwarten.[477] Unter Umständen lassen sich so bisher nicht-erklärbare psychische UAWs der COX-2-Inhibitoren (z.B. Depression bei Celecoxib[478]) über die Interaktion mit dem CB-Rezeptor erklären.

b) **PPAR-γ-Rezeptor und COX:** bestätigte Kreuzreaktivität

Verschiedene COX-1/2 Inhibitoren (z.B. Indometacin, Diclofenac, Ibuprofen) sind im µM-Konzentrationsbereich PPAR-γ Agonisten[479–480]. Hier ist ein therapeutischer Synergismus z.B. in der Tumor-Therapie[481–482] zu erwarten, jedoch resultieren aus Liganden mit dualer Aktivität auch potentielle UAWs (z.B. peptische Ulcera durch COX-Inhibition).

c) **PPAR-γ-Rezeptor und CB-Rezeptor:** bestätigte Kreuzreaktivität

Einige Cannabinoide (z.B. Ajulemic acid) sind im µM-Konzentrationsbereich PPAR-γ Agonisten[483]. Die Differenzierung von Fibroblasten zu Adipozyten oder die antiinflammatorische cannabinoide Wirkung sind auf PPAR-γ-Agonismus zurückzuführen.[483] Einige längerfristige therapeutische Effekte der Cannabinoide lassen sich durch die Aktivierung des PPAR-γ-Rezeptors erklären.[484] Es sind additive oder synergistische positive Effekte durch diese Interaktion zur Therapie von inflammatorischen oder Immunerkrankungen (z.B. Morbus Alzheimer, Morbus Parkinson, Multiple Sklerose) zu erwarten.[485] Neue Indikationen für alte Arzneistoffe sind möglich: Einsatz des CB-1-Antagonisten Rimonabant (ebenfalls experimentell bestätigter PPAR-γ-Agonist[483]) bei übergewichtigen Diabetes Typ 2 Patienten aufgrund typischer PPAR-γ-Effekte auf den Lipid- und Glucose-Stoffwechsel.[486] Jedoch ist bei Rimonabant zu berücksichtigen, dass der Arzneistoff 2008 aufgrund eines gesteigerten Suizidrisikos

und vermehrtem Auftreten von Depressionen vom Markt genommen wurde.[487] Insbesondere bei einer geplanten Langzeittherapie von Diabetes Typ 2 Patienten könnte dies sehr kritisch sein.

B) PDE-5 und Adenosin-A3-Rezeptor:

Zwischen der Phosphodiesterase 5 (Enzym) und dem Adenosin-A3-Rezeptor (GPCR), die von von der Proteinsequenz völlig unterschiedlich sind, wird ebenfalls eine potentielle Kreuzreaktivität gefunden. Bestätigt wird dies auch durch die Literatur.[488–489]

Eine einfache Erklärung für diese strukturelle Ähnlichkeit von einigen Liganden findet man über die endogenen Liganden. Phosphodiesterasen bauen den sekundären Botenstoff zyklisches Guanosin- oder Adenosinmonophosphat (cGMP oder cAMP) zu GMP bzw. AMP ab.[490] Natürlicher Ligand am Adenosin-Rezeptor ist das Purinnucleosid Adenosin, das Bestandteil des cAMP/AMP ist.[491] Entsprechende Inhibitoren bzw. (Ant)agonisten an diesen Targets stellen oftmals strukturelle Analoga dieser natürlichen Liganden dar. Für den PDE-5-Inhibitor Sildenafil, der ein Guanin-ähnliches Struktur-Element enthält, ist beispielsweise in ChEMBL K_i-Wert von 142nM für den Adensosin-A2a-Rezeptor zu finden. Diese Multi-Target-Aktivität kann z.B. zu einer weiteren Verstärkung der vasodilatierenden Wirkung des Silfdenafils führen.

C) XDH und Kinasen

Eine Kreuzreaktivität wurde ebenfalls zwischen der Xanthindehydrogenase und den Kinasen JNK-3 und CDK-5 gefunden. Einige Liganden mit dualer Inhibition konnten in der BindingDB und Literatur gefunden werden, darunter auch einige Flavonoide[492]. Bei Flavonoiden wäre zu prüfen, ob es sich um echte Inhibitoren oder „Assay-Artefakte“ z.B. aufgrund von optischer Interferenz mit dem Assay-System handelt (vgl. Abschnitt 2.2.2).

Eine Erklärung für die Kreuzreaktivität liefern ebenfalls die endogenen Liganden der Targets. Die Xanthindehydrogenase katalysiert die Oxidation des Purins Hypoxanthin zu Xanthin bzw. Xanthin zu Harnsäure.[493] Der bekannteste therapeutisch eingesetzte XDH-Inhibitor Allopurinol stellt ein Xanthin-Analogon dar. Kinasen benötigen ATP zur Phosphorilierung ihrer Substrate. Die meisten Kinase-Inhibitoren binden in der ATP-Bindetasche und weisen z.T. Adenin- oder Purin-ähnliche Struktur-Elemente auf (vgl. Kapitel 20.2.1). Zudem ist für die XDH bekannt, dass sie eine Vielzahl N-haltiger Heterozyklen oxidiert, die ebenfalls häufig Grundstrukturen von Kinase-Inhibitoren darstellen. Eine Inhibition der XDH und JNK-3 kann therapeutisch im Bereich (kardio)vaskulärer oder neurodegenerativer Erkrankungen u.a. durch Reduktion des oxidativen Stresses vorteilhaft sein.[492, 494–495]

D) Aromatase und Androgen/Estrogen-Rezeptor

Die Ähnlichkeit von Aromatase (Enzym) und Androgen- bzw. Estrogen-Rezeptor lässt sich ebenfalls leicht über die endogenen Liganden der Targets bzw. die Substrate und Produkte, der Aromatase erklären. Die Aromatase (CYP2C19) katalysiert die Estrogen-Biosynthese (z.B. Estradiol), wobei als Substrate Androgene wie z.B. Testosteron dienen.[496] Steroidale

Aromatase-Inhibitoren stellen entsprechende Substrat-, Intermediat- oder Produkt-Analoga dar und weisen daher große Ähnlichkeit zu Liganden des Androgen- bzw. Estrogen-Rezeptors auf.[497]

E) DAT/SERT/NET und verschiedene aminerge GPCRs

Die Transporter DAT, NET und SERT sind für die zelluläre bzw. neuronale Wiederaufnahme der Neurotransmitter Dopamin, Noradrenalin und Serotonin nach Ausschüttung verantwortlich. Diese Transporter haben somit die gleichen endogenen Liganden wie aminerge GPCRs. Für zahlreiche Antidepressiva und Neuroleptika ist die Wirkung auf die gleichzeitige Modulation verschiedener aminerger GPCRs und der Monoamin-Transporter zurückzuführen.[307, 463] Die vielfältigen UAWs der trizyklischen Antidepressiva (z.B. Imipramin) im Vergleich zu den selektiven Serotonin-Reuptake-Inhibitoren wie Fluoxetin sind durch die zusätzliche Bindung an verschiedenste aminerge GPCRs (z.B. 5-HT_{2A}, 5-HT_{2C}, 5-HT_6, Alpha-1, Muskarin und Histamin-Rezeptoren) bedingt.[463]

F) MAO und aminerge GPCRs

MAO katalysiert die Desaminierung von Monoaminen (z.B. MOA-A hat physiologisch die Funktion die biogenen Aminen Serotonin und Noradrenalin abzubauen, MAO-B baut Dopamin ab). Diese Monoamine stellen auch endogene Liganden an aminergen GPCRs dar. Eine Ähnlichkeit zwischen diesen Targets ist in Anbetracht gleicher endogener Liganden sinnvoll.

G) 3-HT_3-Rezeptor und aminerge GPCRs

Da Serotonin endogener Ligand sowohl an ionotropen (5-HT_3) als auch an metabotropen Serotonin-Rezeptoren (z.B. 5-HT_{2B} oder 5-HT_{1A}) ist, ist eine Kreuzreaktivität zu erwarten. Für Alosetron ist eine Modulation des 5-HT_{2B} (IC_{50} =18 nM) experimentell bestätigt.[348] Für andere Setrone ist ebenfalls eine Modulation des 5-HT_{1A} bekannt.[470] Weitere Kreuzreaktivitäten mit anderen aminergen GPCRs erscheinen wahrscheinlich.

H) AChE und Muskarin-Rezeptoren

Die Acetylcholinesterase katalysiert physiologisch die Esterhydrolyse des Neurotransmitters Acetylcholin. Das natürliche Substrat dieses Enzyms ist ebenfalls der endogene Ligand an muskarinischen Acetylcholin-Rezeptoren (GPCRs). Eine Kreuzreaktivität ist somit zu erwarten. Experimentell ist für verschiedene AChE-Inhibitoren (z.B. Physiostigmin) eine zusätzliche Affinität für verschiedene Muskarin-Rezeptoren (z.B. M_1, M_2, M_4) bestätigt.[498–499] Diese Kreuzreaktivität kann mit verstärkten muskarinischen UAWs (z.B. GIT-Störungen) einhergehen. Aufgrund der wichtigen Rolle des cholinergen Systems bei mnestischen Störungen kann man diese Multitarget-Aktivität auch therapeutisch zur Verbesserung der Gedächnisleistung (z.B. von Schizophrenie-Patienten) nutzen.[500]

I) PG-E-Synthase/COX-1 und PG-D2-Rezeptor

Die PG-E-Synthase katalysiert die Umwandlung des Prostaglandin-H2 in Prostaglandin-E. Der Prostaglandin-D2-Rezeptor (GPCR) bindet das zum Prostaglandin-E strukturell sehr ähnliche Prostaglandin-D2. Die Cyclooxygenase katalysiert hingegen die Bildung von Prostaglandin-H2 aus Arachidonsäure via Prostaglandin-G2. Die gefundene Ähnlichkeit ist aufgrund der ähnlichen Liganden bzw. Substrate oder Produkte und die gemeinsame Involvierung in den Prostaglandin-Stoffwechsel leicht zu erklären. Da entsprechende Inhibitoren/Antagonisten zumeist Analoga der Liganden/Substrate/Produkte darstellen, ist eine klinisch relevante Kreuzreaktivität aufgrund überlappendender pharmakophorer Eigenschaften (vgl. ähnliche Multi-Target-Liganden bei MORPHY et al.[470]) wahrscheinlich.

2.) Entwicklung dualer Liganden

Beispiel H_3-Rezeptor und BChE

Einige gefundene Kreuzreaktivitäten, die sich auch durch Literaturrecherchen belegen lassen, bilden auch Strategien in der Arzneistoff-Entwicklung ab. Das heißt, dass gezielt duale Liganden oder Multi-Liganden zur effektiveren Therapie einer komplexen Erkrankung, für die mehrere assoziierte Targets identifiziert werden können (vgl. Kapitel 1), entwickelt werden. Ein Beispiel hierfür stellen einige H_3-Rezeptor-Antagonisten, die zugleich im nanomolaren Konzentrationsbereich auch Inhibitoren der Butyrylcholinesterase darstellen.[501] Wichtige gemeinsame pharmakophore Eigenschaften scheinen hierbei ein positiv ionisierbares Zentrum in einem bestimmten Abstand zu aromatischen Ringen (vgl. Abbildung 23.10) und ggf. noch einer positiven Ladung zu sein. Durch gleichzeitige Modulierung beider Targets ist mit synergistischen Effekten in der Therapie des Morbus Alzheimer zu rechnen[501], jedoch sind auch cholinerge UAWs sehr wahrscheinlich.

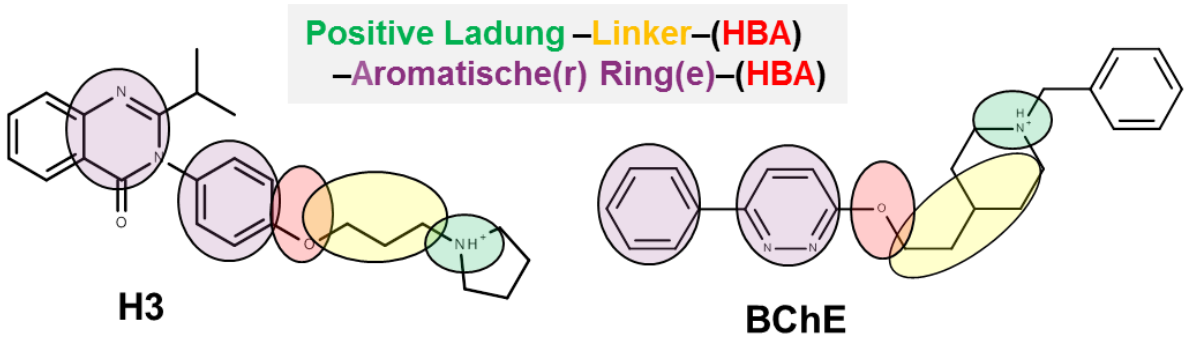


Abbildung 23.10. Typische gemeinsame Merkmale von Liganden am H3-Rezeptor (H_3, GPCR) und der Butyrylcholinesterase (BChE, Enzym) abgeleitet anhand der Menge an gemeinsamen MCSs.

Weitere Beispiele für entwickelte Multitarget-Liganden

MORPHY et al. geben eine Auflistung von Target-Kombinationen in 92 entwickelten Multi-Target-Arzneistoffen.[470, 502] Viele dieser Kombinationen spiegeln sich auch im inSARa-basierten Ähnlichkeits-Netzwerk wider (z.B. MMP/Cathepsin, TS/DHFR, COX-2/5-LOX, Opioid-μ/κ/δ, SERT/DAT/NET, SERT/α_2, SERT/5-HT_{1A}, SERT/5-$HT_{1B/D}$, H_1/H_3, Adenosin-A1/3).

3.) Entwicklungsgeschichtliche strukturelle Ähnlichkeit

Beispiel Estrogen und COX-2

Eine weitere Ursache für eine hohe strukturelle Ähnlichkeit bzw. ähnliche pharmakophore Eigenschaften der Liganden zweier nicht-verwandter Targets können auch auf die Entwicklungsgeschichte einer Substanzklasse zurückzuführen sein. Zum Beispiel werden die ursprünglichen UAWs einer Substanzklasse in manchen Fällen zu therapeutischen Zwecken genutzt werden. Bei der Optimierung der Off-Target-Aktivität zur Hauptwirkung bleiben meist einige strukturelle Merkmale der ursprünglichen Substanzklasse erhalten, die u.U. bei Analysen dieser Art als potentielle Kreuzähnlichkeit in Erscheinung tritt.

So lassen sich die Ursprünge der COX-2-Inhibitoren Celecoxib und Rofecoxib auf die Entwicklung nicht-steroidaler Estrogene (z.B. Chlortrianisen) zurückführen, woraus sich über Umwegen der Estrogen-Rezeptor-Modulator Raloxifen ableitet (vgl. Abbildung 23.11).[503] Ein Indolanalogon hiervon zeigte unerwartete antiinflammatorische Wirkung und war der Ausgangspunkt für die Entwicklung eines dem Celecoxib strukturell sehr ähnlichen COX-Inhibitors.[503]

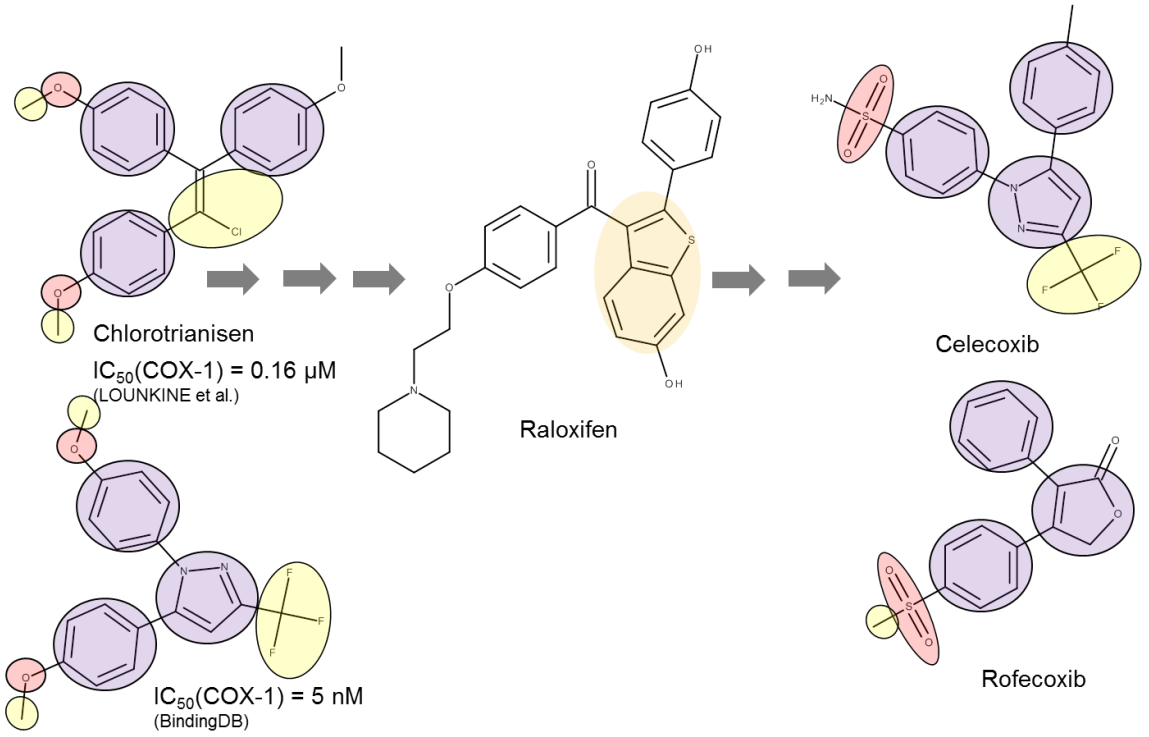


Abbildung 23.11. Entwicklungsgeschichtliche[503] strukturelle Gemeinsamkeiten als Erklärung für Kreuzreaktivität zwischen Cyclooxygenase (COX) und Estrogen-Rezeptor. Gemeinsame pharmakophore Eigenschaften farbig markiert (Aromat = lila, lipophil = gelb, HBA = rot). Die angegebenen Bioaktivitätsdaten stammen von LOUNKINE et al.[348] und aus der BindingDB.

Dass aber zwischen Liganden an der COX und dem Estrogen-Rezeptor tatsächlich Kreuzreaktivitäten bestehen, bestätigt die positive experimentelle Testung des Estrogen-Agonisten Chlorotrianisen an der COX-1 (IC_{50} = 0.16µM) von LOUNKINE et al.[348], die eine Erklärung beispielsweise für die abdominellen UAWs des Estrogen-Agonisten liefert.

Deutliche strukturelle Ähnlichkeiten (auch im Vergleich mit Abbildung 23.9) lassen sich zwischen den Estrogen-Rezeptor-Liganden und den COX-Liganden in Abbildung 23.11 erkennen. Beim Vergleich der pharmakophoren Eigenschaften des Chlorotrianisens mit denen eines potenten COX-1-Inhibitors aus der BindingDB erscheint die von LOUNKINE et al.[348] beschriebene Off-Target-Aktivität weniger überraschend.

4.) Sonderfall: Der σ-Rezeptor als „pharmakologisches Enigma“[504]

Der σ-Rezeptor stellt einen Sonderfall unter den analysierten Targets dar. Er weist eine Ähnlichkeit zu einer unglaublichen Vielzahl verschiedener Targets auf: Ionenkanäle (nAChR, Glutamat-NMDA), Transporter (SERT, NET, DAT), peptidische GPCRs (Opioid-µ, -κ, -δ, NK-1), Enzyme (AChE, BChE) und aminerge GPCRs (5-$HT_{1a, 1b, 2a, 2b, 2c, 6}$, M_1-M_4, H_1/$_3$, D_1-D_4, β_2, α_{1A}, α_{2C}). Schaut man sich die Liste bisher beschriebener Liganden an dieser Zielstruktur an, so ist eine ebenso große Variabilität festzustellen, die eine sinnvolle Erklärung für die hergestellten Kreuzreaktivitäten liefert. Eine Zusammenstellung bekannter Arzneistoffe oder missbräuchlich verwendeter Psychostimulantien, die als σ-Rezeptor-Liganden (Agonisten und Antagonisten) beschrieben sind, und mit ihnen assoziierte Targets liefert Tabelle 23.1.

Das gemeinsame pharmakophore Motiv aller Liganden ist relativ unspezifisch (positiv ionisierbares Zentrum über Linker verknüpft mit einem Aromaten[505]), was eine große strukturelle Diversität erklärt.

Der σ-Rezeptor hat lange Zeit als „rätselhaftes Protein" gegolten, weil er sich nicht in die üblichen Klassifikationsschemata einteilen ließ (daher Klassifikation „Orphan" in Tabelle 17.2).[506] Erst fehlklassifiziert als Opioid-Rezeptor, wurde 2007 erkannt, dass er ein einzigartiges ligand-gesteuertes Chaperon des Endoplasmatischen Retikulums darstellt, das modulierende Signalfunktion zwischen den Zellorganellen zu haben scheint.[506–508] Der σ-Rezeptor spielt u.a. eine wichtige Rolle in der Pathophysiologie psychischer Erkrankungen, der Schmerzwahrnehmung, der Gedächtnisfunktion und ist mit neuroprotektiven Effekten (Morbus Alzheimer, Morbus Parkinson) assoziiert.[509] Für die Entwicklung von neuen Arzneistoffen ist er somit als On-/Off-Target von großer Bedeutung.

Im Bereich der psychischen Erkrankungen ist nicht nur eine Vielzahl verschiedener Targets involviert, sondern Arzneistoffe derselben Klasse zeigen in vielen Fällen unterwartete Unterschiede in der therapeutischen Wirksamkeit. Biologische Aktivität an weiteren, ggf. noch unbekannten Zielstrukturen wie z.B. dem σ-Rezeptor erklären oftmals diese variierenden pharmakologischen Profile. Zusätzliche Affinität für den σ-Rezeptor stellt beispielsweise eine mögliche Erklärung für die bisher nicht erklärbare unterschiedliche therapeutische Wirksamkeit von bestimmten Antidepressiva dar (z.B. Überlegenheit von der Sertralin und Fluvoxamin gegenüber Paroxetin, obwohl alle SSRIs darstellen).[504] Ebenso scheint die Affinität für den σ-Rezeptor eine wichtige Rolle bei der Erklärung von schwer zu therapierbaren Symptomen nach Langzeit-Missbrauch von Psychostimulantien wie Cocain oder Methamphetamin zu spielen.[510]

Tabelle 23.1. Übersicht über bekannte Arzneistoffe, die Liganden am σ-Rezeptor sind, und einige Targets, an denen sie ebenfalls biologische Aktivität zeigen.

Ligand	**Assozierte Targets**
Opiode (Morphin, Pentazocin etc.)[504, 509]	Opioid-μ, -κ, -δ
Dextromethorphan[504]	Glutamat-NMDA, SERT, NET, (Opioid-μ, -κ, -δ, nAChR)
SSRIs (Sertralin, Fluoxetin, Fluvoxamin)[504]	SERT
Trizyklische Antidepressiva (Imipramin[508], Opipramol[511])	SERT, NET (verschiedene aminerge GPCRs)
Neuroleptika (Haloperidol[504], Chlorpromazin[509])	v.a. D_2 und 5-HT_{2a} (5-HT_{2c}, $D_{1\text{-}5}$, α_{A1}, H_1, $M_{1/4}$ etc.)
Donepezil[512]	AChE (BChE, M, nAChR)
Phencyclidin (PCP)[508]	Glutamat-NMDA
Cocain[508]	SERT, DAT, NET, MAO
Methamphetamin[508]	DAT, SERT, NET
MDMA[509]	SERT, DAT, NET
Chlorpheniramin[508]	H_1 (weitere aminerge GPCRs)
Amantadin, Memantine[509]	Glutamat-NMDA

Kurz-Zusammenfassung der Recherchen zu weiteren potentiellen Kreuzreaktivitäten

Tabelle 23.2 fasst die Recherche-Ergebnisse zu weiteren Kreuzreaktivitäten zusammen. Dabei lässt sich feststellen, dass sich die meisten Kreuzreaktivitäten durch Literatur-Recherche bestätigen lassen. Bei Target-Paaren, für die keine experimentelle Bestätigung gefunden werden konnte, ist trotzdem zumeist ein gemeinsamer Signalweg in der Literatur beschrieben (z.B. β-Adrenorezeptor und TR-β).

Cathepsin-B ist ein Target, für das Ähnlichkeiten zu verschiedenen Zielstrukturen anderer Target-Klassen auf Basis der inSARa-Netzwerke gefunden werden (z.B. BACE-1, ABL-1, FXa, Melanocortin-3). Experimentell lässt sich dies jedoch nur im Fall der Beta-Sekretase-1 bestätigen. Möglicherweise ist die Ähnlichkeit zu den anderen Targets auf die unspezifische, peptidische Struktur der jeweiligen Inhibitoren zurückzuführen (vgl. ABL-1/CatB in Abschnitt 23.1.5), die oftmals zu großen RGs und folglich auch großen, unspezifischen MCSs führt.

Tabelle 23.2. Zusammenfassung der Beurteilung weiterer potentieller inSARa-Netzwerk-Kreuzreaktivitäten. Abkürzung: (+) = experimentelle Bestätigung, (+/-) = keine experimentelle Bestätigung, aber Kreuzreaktivität möglich, (-) Kreuzreaktivität unwahrscheinlich

Target A	Target B	Bestätigung der Kreuz-reaktivität?	Quelle/weitere Informationen
MAO	Estrogen	(+)	ULUS et al.[513]: Estradiolbenzoat = schwache, reversible MAO-Inhibition
MAO-A/B	CB-1/2	(+)	FISAR[514]: Inhibition der MAO-A/B durch verschiedenen Cannabinoide (z.B. THC, Anandamid) im niedrigen µM-Bereich → Nicht-CB-Rezeptor-vermittelte Modulation der Monoamin-Neurotransmission im Gehirn → Erklärung für stimmungs- und emotionsbeeinflussende Effekte von Cannabinoiden[515]
5-LOX	HIV-Integrase	(+)	BAILLY et al.[516]: antioxidative HIV-Integrase-Inhibitoren
COX	MAO	(+)	NAKATANI et al.[517]/OSHISHI et al.[518]: Xanthone-Derivate sind COX-Inhibitoren und reversible MAO-Inhibitoren
JNK-3	Estrogen-β	(+/-)	Keine experimentelle Bestätigung gefunden, aber Tumor-suppressiver Effekt hoher Dosis von Estrogen auf JNK-Aktivierung zurückzuführen[519–520]; Inhibitor anderer Kinase ist partieller Antagonist am Estrogen-Rezeptor[521]
CCR5	M_2	(+)	LIN et al.[349]: Mehr als 30 Antagonisten aus verschiedenen Liganden-Serien zeigen Aktivität an beiden Targets
Somato-statin-5	H_1	(+)	LIN et al.[349]: Ligandensätze weisen große Gemeinsamkeiten auf
TR-β	β_2	(+/-)	Keine experimentelle Bestätigung gefunden, aber Thyroid-Hormone haben vielfältige direkte (Erhöhung der Bindungsaffinität) und indirekte Effekte (Steigerung der Rezeptor-Proteinbiosynthese) auf β-Adrenorezeptor[522], β-adrenerge Effekte bei Hyperthyreose (z.B. Tachykardie, Tremor, Glykogenolyse)[523]
Melano-cortin-3/	CatL/B	(-)	Keine Bestätigung gefunden, Ähnlichkeit möglicherweise auf peptidische Struktur der jeweiligen Inhibitoren zurückzuführen (vgl. ABL-1/CatB in Abschnitt 23.1.5)
CatB	ABL-1	(-)	Keine Bestätigung gefunden, für Details vgl. Abschnitt 23.1.5
CatB	BACE-1	(+)	HOOK et al.[524]: CatB-Inhibitoren haben in BACE-1-Assay ebenfalls IC_{50} niedrig-nM-Bereich → Alzheimer Therapie
CatB	Trypsin/ FXa	(-)	Keine Bestätigung gefunden (vgl. Melanocortin-3/CatB)
COX-2	P38	(+)	Für Details vgl. Abschnitt 23.1.5
Orexin-2	MAO	(+/-)	Keine experimentelle Bestätigung gefunden, aber es ist bestätigt, dass verhaltensbeeinflussende Effekte des Orexin-Rezeptors partiell MAO(-A) vermittelt sind[525]
Orexin-1/2	Estrogen-β	(+/-)	Keine experimentelle Bestätigung gefunden, aber der Orexin-1/2 reguliert LH-Freisetzung[526] und gemeinsame Beeinflussung des weiblichen Verhaltens durch Estrogen und Orexin bestätigt[527]
Orexin-1	H_3	(+/-)	Keine experimentelle Bestätigung gefunden, aber der Orexin-1/2-Rezeptor spielt ebenso wie der $H_{1/3}$-Rezeptor eine wichtige Rolle bei der Regulation von Wachheit[528–529] → Targets zur Therapie der Narkolepsie[530]

23.1.5. Beurteilung der Bedeutsamkeit von potentiellen Ähnlichkeiten/Kreuzreaktivitäten mittels inSARa-Netzwerke

Der Vorteil dieses Ansatzes im Vergleich zu anderen Ligand-basierten Analysen von Ähnlichkeiten bzw. Kreuzreaktivitäten zwischen Zielstrukturen, die in der Regel auf Fingerprint-Ähnlichkeit beruhen (vgl. SEA, Abschnitt 8.2), ist die direkte Interpretierbarkeit der berechneten Ähnlichkeiten. Bei dem hier verwendeten MCS-basierten Ansatz können sofort die Merkmale, die für die Gemeinsamkeiten verantwortlich sind, identifiziert werden. Durch Betrachten dieser gemeinsamen MCSs zweier Zielstrukturen mit den zugehörigen Liganden in den inSARa-Netzwerken kann der medizinische Chemiker sehr schnell entscheiden, ob die gezeigten Beziehungen von Relevanz für Kreuzreaktivitäten sind. Dies soll im Folgenden an zwei unterschiedlichen Beispielen veranschaulicht werden.

Beispiel 1: Cathepsin-B und ABL-1

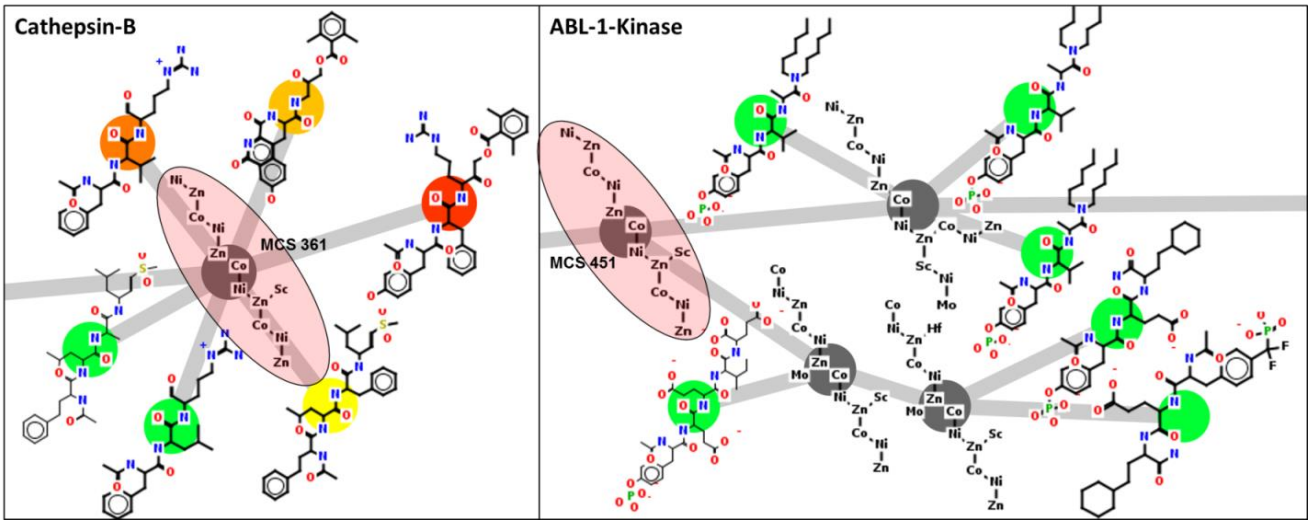


Abbildung 23.12. Beurteilung potentieller Kreuzreaktiviäten mit Hilfe von inSARa-Netzwerken am Beispiel von Cathepsin-B und der ABL-1-Kinase. Der Knoten des größten MCS der Liste der gemeinsamen MCSs beider Targets ist rot markiert und die zugehörigen Moleküle sind an den Molekülknoten abgebildet (Bioaktivität IC_{50}: 10µM > grün > 1µM > gelb > 10nM > rot)

Zwischen der Cystein-Protease Cathepsin-B und die ABL-1-Kinase wurde auch eine unerwartete Ähnlichkeit festgestellt. Obwohl Überexpression bzw. Dysregulation beider Enzyme eine wichtige Rolle in der Progression verschiedener Tumorerkrankungen spielt[531-532], konnte kein experimenteller Nachweis für diese potentielle Kreuzreaktivität gefunden werden.

Bei Betrachten der Liste gemeinsamer MCSs kann festgestellt werden, dass zusätzlich zu verschiedenen kleineren MCSs einige sehr große MCSs auftauchen. Bei den großen MCSs lässt sich feststellen, dass sie eher unspezifischer Natur sind. Verschiedene Abfolgen des Motives HBA (=Ni) – HBD (=Co) – Linker (=Zn) und ein aromatischer Ring (=Sc) sind Bestandteile dieser MCSs. Dies ist typisch für peptidische Strukturen, wobei Co und Ni zumeist die Amidbindung repräsentieren.

Da sehr große MCSs eine hohe Wahrscheinlichkeit aufweisen, dass die zugehörigen Moleküle auch sehr ähnlich sind, ist es zur Beurteilung von potentiellen Kreuzreaktivitäten sinnvoll sich die größten gemeinsamen MCSs in den inSARa-Netzwerken anzeigen zu lassen. In Abbildung 23.12 ist der größte MCS aus der Liste der gemeinsamen MCSs von Cathepsin-B und der ABL-1-Kinase in inSARa-Netzwerken der beiden Targets dargestellt (Knoten 361 und 451). Folgende Dinge lassen dabei feststellen:

- Im Cathepsin-B-Netzwerk stellt der MCS-Knoten 361 einen SAR Hotspot dar. Wie zu erwarten scheint dieser unspezifische MCS selber nicht die Bioaktivität am Enzym zu bestimmen, sondern weitere pharmakophore Eigenschaften scheinen dafür entscheidend zu sein.
- Im Netzwerk der ABL-1-Kinase stellt MCS-Knoten 451 einen intermediären Knoten dar. Wie in der Abbildung zu erkennen, gibt es weitere größere MCS-Knoten, an denen die zugehörigen Moleküle stecken bleiben. Anhand der Farbe der Molekül-Knoten kann man erkennen, dass die Moleküle nur sehr schwach aktiv (IC_{50} zwischen 1 und 10µM) sind. Potente Kinasen-Inhibitoren weisen normalerweise eine Bioaktivität im nanomolaren bis subnanomolaren Konzentrationsbereich auf. Zudem ist die peptidische Struktur sehr ungewöhnlich für Kinase-Inhibitoren. Auffällig bei den Molekülen ist zudem die negativ ionisierte Phosphono- oder Phosphono-difluoromethyl-Gruppe (Phosphatase-stabiles Analogon mit verbesserter Membranpermeation[533]) am Phenylring. Dank des direkten Verweises der BindingDB auf die Primärquellen[533–534] lässt sich dies erklären: Bei diesen Peptidomimetika handelt es sich um Phospho-Tyrosin-Mimetika. Diese binden im Gegensatz zur Mehrheit der Kinase-Inhibitoren, die kompetitiv in der Kinase-Domäne in der ATP-Bindestelle binden[535], an die src-Homologie (SH2) Domäne. Diese ist bei einer Vielzahl von cytoplasmatischen Proteinen, die in die intrazelluläre Signaltransduktion involviert sind, konserviert und vermittelt essentielle Protein-Protein-Interaktionen.[536]

Nach dieser Analyse lässt sich somit leicht feststellen, dass eine Kreuzreaktivität unwahrscheinlich ist. Zudem können mögliche Fehlerquellen erkannt bzw. Optimierungsmöglichkeiten für die Methode abgeleitet werden:

- Die Codierung von peptidischen Strukturen führt häufig zu sehr großen RGs. Der MCS dieser RGs repräsentiert häufig das unspezifische peptidische Rückrat und führt zu relativ großen MCSs. In dem gezeigten Beispiel bestand der MCS aus 12 Pseudoatomen. Bei kleinen Molekülen mit nicht-pepditischer Struktur weist ein MCS dieser Größe, wie eine Vielzahl von Analysen gezeigt haben, eine hohe Spezifität auf und codiert mit hoher Wahrscheinlichkeit strukturell sehr ähnliche Moleküle. Bei normalen Arzneistoff-Molekülen wäre hier daher die Wahrscheinlichkeit für eine Kreuzreaktivität hoch. Aus diesem Grund wird die MCS-Größe bei der Ähnlichkeitsberechnung zur Gewichtung verwendet. Wie dieses Beispiel zeigt, stellt dies aber auch eine Fehlerquelle dar. So ist es möglich, dass einzelne sehr große MCSs (z.B. verursacht durch verschiedenartige peptidische Strukturen) zu einer unbedeutenden Ähnlichkeitsbeziehung führen. Zur Optimierung könnte ein Maximalgewicht eingeführt werden, das ab einer bestimmten MCS-Größe Verwendung findet, um Übergewichtung durch einzelne MCSs zu vermeiden.

- Es ist auch festzustellen, dass die RG-Codierung für die Analyse bestimmter makromolekularer Moleküle weniger gut geeignet ist. Wie von STIEFL et al. bei ihrem

ErG-Ansatz bereits festgestellt, führen peptidische Moleküle zu einem „Information-Überfluss“, bei dem die wichtigen pharmakophoren Eigenschaften, die durch die Seitenketten codiert werden, in der Vielzahl an nicht-interagierenden Linker bzw. Amid-Atomen im RG verloren gehen.[239] Zur Optimierung wäre z.B. speziell für die Analyse von peptidischen Molekülen eine alternative RG-Codierung denkbar. Dasselbe gilt für die Analyse von Naturstoffe, die ebenfalls aufgrund z.B. des überproportionalen Vorkommens an funktionellen Gruppen (z.B. Hydroxylgruppen in Glykosiden), zu sehr großen RGs und ggf. auch (unspezifischen) MCSs führen könnten. Es ist jedoch zu beachten, dass bei dieser Analyse Moleküle dieser Art meist schon über den eingeführten Schwellenwert bei der Molaren Masse ausgeschlossen werden. Kleine Peptide sind jedoch in den Datensätzen zu finden.

- Neben einer veränderten Gewichtung könnte man bei der Analyse den Mindest-Aktivitäts-Schwellenwert auf <1µM senken, um die Relevanz von potentiellen Kreuzreaktivitäten zu erhöhen. Die gezeigten schwach potenten SH2-Inhibitoren wären dann beispielsweise nicht Bestandteil des ABL-1-Kinase Datensatzes gewesen und wären folglich nicht in der Analyse berücksichtigt worden. Bei Kinase-spezifischen Analysen (z.B. SUTHERLAND et al.[344]) werden meist nur Moleküle mit nanomolarer Bioaktivität berücksichtigt. Ggf. könnte man speziell bei Datensätzen von Targets den Bioaktivitäts-Schwellenwert anpassen, bei denen hohe Affinitäten für relevante Interaktionen vorausgesetzt werden.

Beispiel 2: P38 und COX-2

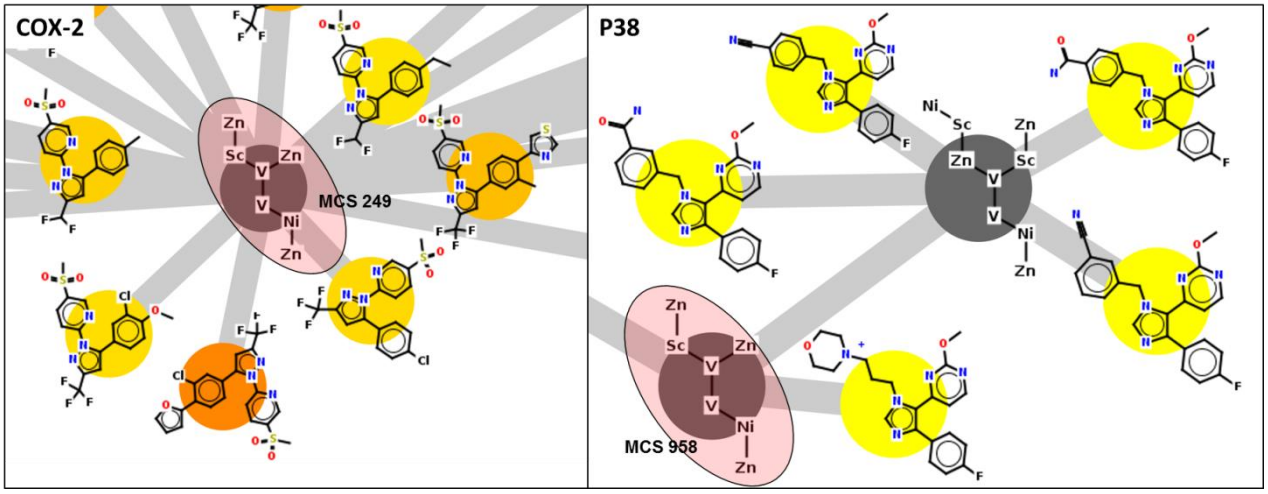


Abbildung 23.13. Beurteilung potentieller Kreuzreaktiviäten mit Hilfe von inSARa-Netzwerken am Beispiel der Cyclooxygenase-2 (COX-2) und der MAP-Kinase p38 alpha (P38). Der Knoten eines großen MCS der Liste der gemeinsamen MCSs beider Targets ist rot markiert und einige zugehörige Moleküle sind an den Molekülknoten abgebildet (Bioaktivität IC_{50}: 10µM > grün > 1µM > gelb > 10nM > rot)

Als zweites Beispiel soll die Beurteilung der Kreuzreaktivität zwischen der Cyclooxygenase-2 und der MAP-Kinase p38 alpha mittels inSARa-Netzwerk veranschaulicht werden. In Abbildung 23.13 ist hierfür ein großer gemeinsamer MCS in den jeweiligen Netzwerken gezeigt. Betrachtet man die zugehörigen Moleküle, so lassen sich deutliche strukturelle

Ähnlichkeiten erkennen: 3 (Hetero-)Aromaten mit z.T. HBA-Funktionen und hydrophobe Gruppen. Auch ein Vergleich der Gesamtmenge an gemeinsamen MCSs zeigt, dass es sich hierbei um typische gemeinsame pharmakophore Eigenschaften handelt (vgl. Zusammenfassung in Abbildung 23.14). Dies ist ähnlich zu dem „Mickey Maus“ Motiv (vgl. Abbildung 5.1 sowie Abschnitt 20.2.2). Diese pharmakophoren Eigenschaften sind auch typisch für Pharmakophor-Modelle dieser beiden Targets[454, 537]. Eine Kreuzreaktivität zwischen einigen COX-2 und P38-Liganden erscheint somit wahrscheinlich.

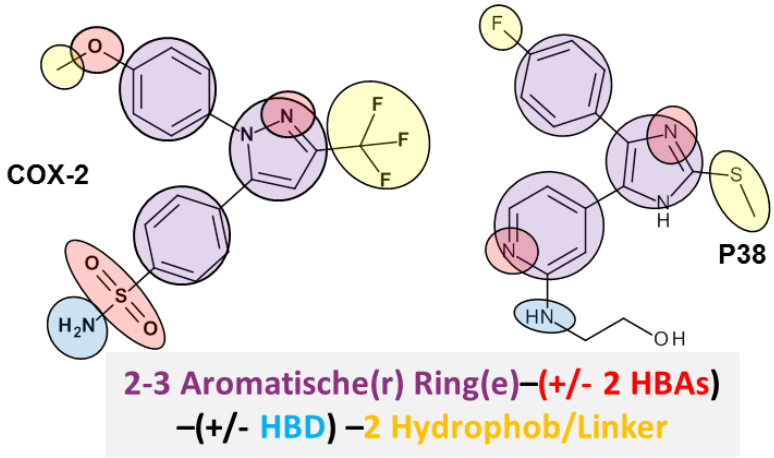


Abbildung 23.14. Typische gemeinsame Merkmale von Liganden an der Cyclooxygenase-2 (COX-2, Oxidoreduktase) und der MAP-Kinase p38 alpha (P38, Kinase) abgeleitet anhand der Menge an gemeinsamen MCSs.

Für einige Pyridinylimidazol-Derivate (z.B: SB203580, vgl. Abbildung 23.15) mit niedrig-nanomolarer Bioaktivität an P38 ist eine direkte Inhibition der COX(-2) in der Literatur beschrieben.[538–540] Docking und COMFA-Studien des COX-2-Inhibitors Celecoxib (vgl. Abbildung 20.22), der strukturelle Ähnlichkeiten zu SB203580 aufweist, zeigen, dass eine Affinität (COMFA-Vorhersage: IC_{50}(P38) = 810nM) für die P38-Kinase zu erwarten ist.[541] Das Docking zeigt, dass die entscheidenden P38-Bindetaschen-Aminosäuren (Met-109 und Lys-53, vgl. Abbildung 23.15) in der besten Docking-Pose ebenfalls Interaktionen mit Celecoxib eingehen.[541] Eine experimentelle Bestätigung der biologischen Aktivität von Celecoxib an der P38-Kinase fehlt jedoch.

Überlappende pharmakophore Eigenschaften beider Targets (vgl. Abbildung 23.14, Abbildung 23.15 und Abbildung 20.22) sind für das Zustandekommen dieser Kreuzreaktivität verantwortlich. Eine zusätzliche Inhibition der COX-2 durch SB203580 ist eine weitere Erklärung für die antiinflammatorische in-vivo-Wirkung des Moleküls[542]. Neben einem Synergismus in der Therapie von entzündlichen Erkrankungen (z.B. Arthritis) durch diese Multi-Target-Aktivität, ist auch in der Tumor-Therapie mit Vorteilen durch duale Inhibitoren zu rechnen.

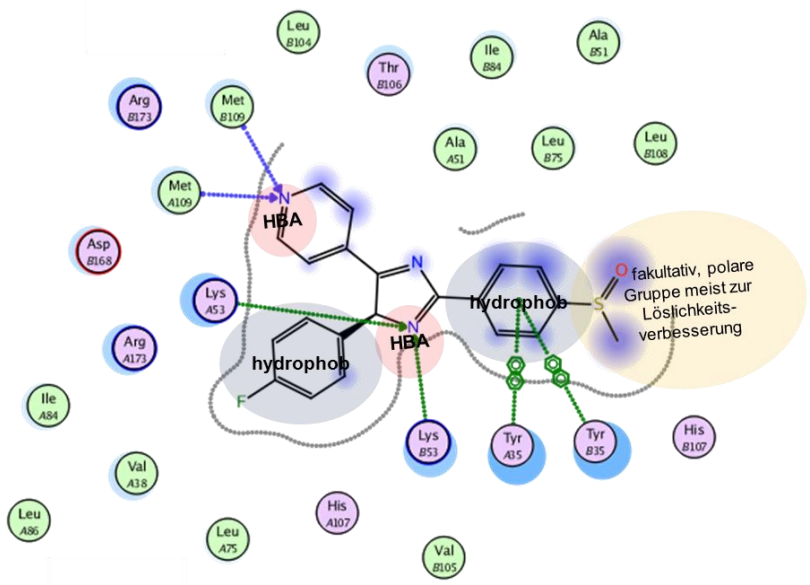


Abbildung 23.15. Affinitätsentscheidende pharmakophore Eigenschaften des P38-Inhibitors SB203580 (PDB-Code: 1A9U; P38) und weiterer Pyridinylimidazol-Derivate (z.B. PDB-Code 1BL6, 1BL7, 1BMK; P38) nach TONG et al.[543] und WILSON et al.[544] (Ligand-Interaktions-Diagramm erstellt mit MOE[188]).

23.1.6. Analyse auf Basis der gesamten MCS-Menge

Im Folgenden werden die Ergebnisse der Analyse auf Basis der gesamten MCS-Menge zusammengefasst. Die gesamte MCS-Menge enthält alle einzigartigen MCSs, die aus dem paarweisen Vergleich aller Datensatz-Moleküle resultieren (vgl. Abschnitt 17.4).

In Abbildung 23.16 sind die Ähnlichkeitswerte des paarweisen Target-Vergleichs für die inSARa-basierte Analyse und die Analyse auf Basis der gesamten MCS-Menge gegeneinander aufgetragen. Die Korrelation (Spearman-Rang-Korrelationskoeffizient) für die Ähnlichkeitswerte auf Basis des gewichteten inSARa-TSim beträgt 0.69. Eine deutliche positive Korrelation ist zu erwarten, da die inSARa-MCSs eine Submenge der Gesamtmenge darstellen. Umso repräsentativer diese Submenge für die Gesamtmenge ist, desto stärker sollte die Korrelation beider Methoden sein.

Der Vergleich mit der Winkelhalbierenden in Abbildung 23.16 zeigt, dass beim Target-Vergleich auf Basis der gesamten MCS-Menge eine erhöhte Ähnlichkeit im Vergleich zur Analyse basierend auf den MCSs aus den jeweiligen inSARa-Netzwerken gefunden wird. Betrachtet man zusätzlich die resultierende Ähnlichkeitskarte (vgl. Abbildung 23.18) so stellt man fest, dass die Werte für die unbedeutende Basis-Ähnlichkeit zwischen Targets („Untergrundrauschen“) deutlich erhöht sind. Dies wird auch durch eine Erhöhung des Medians aller Werte auf 2.81 (zum Vergleich inSARa: 0.22) und des arithmetischen Mittels auf 3.40 (zum Vergleich inSARa: 0.93) deutlich. Folglich ist es bei einer Analyse auf Basis der gesamten MCS-Menge für die Erstellung eines Schwellenwert-Netzwerkes bzw. zur Bewertung von Kreuzreaktivitäten notwendig den Schwellenwert entsprechend zu erhöhen.

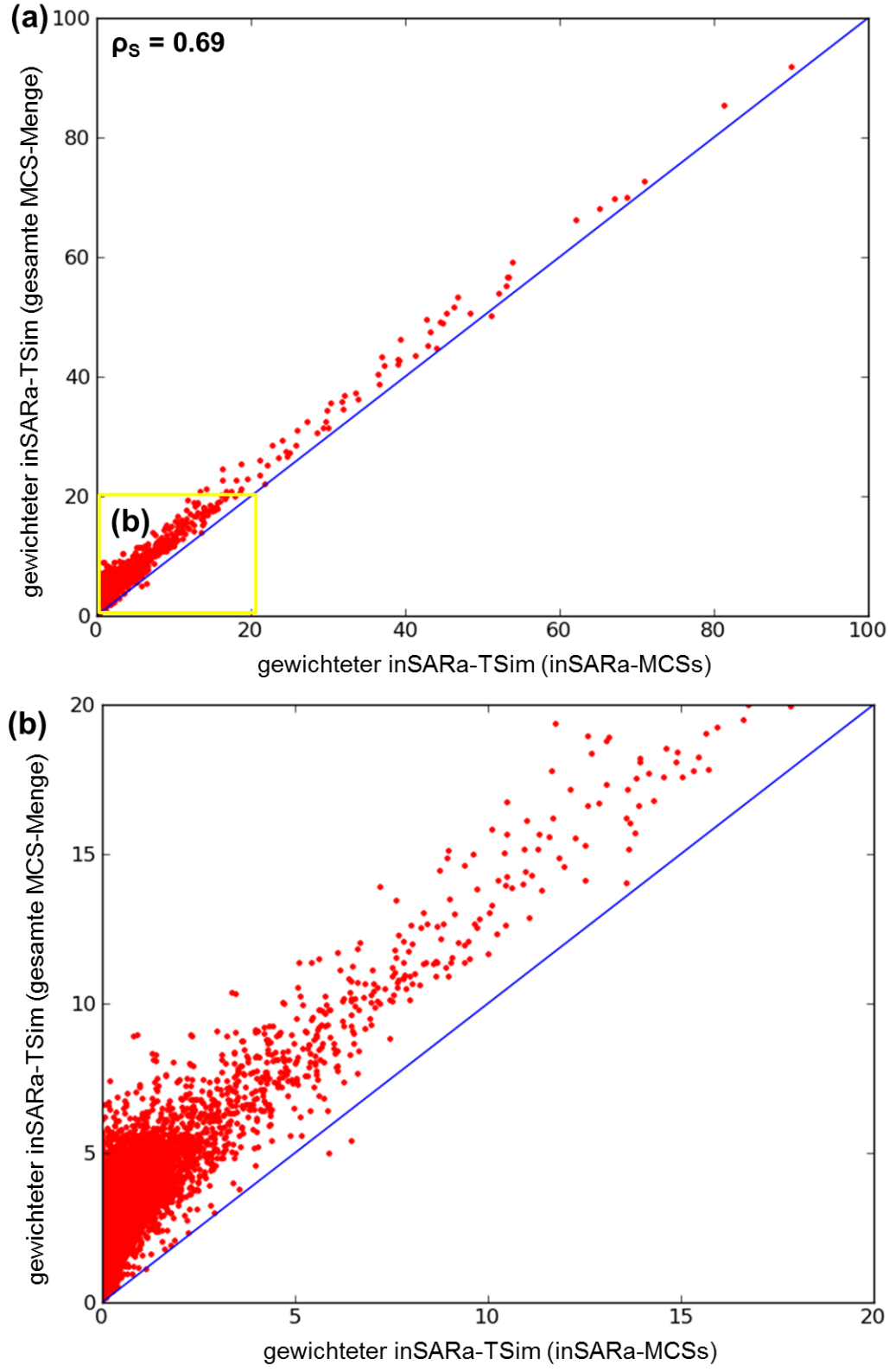


Abbildung 23.16. Korrelation (Spearman-Rangkorrelationskoeffizient ρ_S) der paarweisen Ähnlichkeitswerte aller Targets (gewichteter inSARa-TSim, Eigenvergleiche nicht berücksichtigt) bei Verwendung der inSARa-MCSs oder der gesamten MCS-Menge eines Targets. (b) stellt eine Vergrößerung des gelb markierten Bereiches aus (a) dar.

Da letztendlich die Kanten im resultierenden Schwellenwert-Netzwerk entscheidend für die Auswertung der Analyse sind, wurde zusätzlich untersucht, wie sich die Zahl der Kanten im Schwellenwert-Netzwerk in Abhängigkeit des Ähnlichkeits-Schwellenwertes verändert (vgl. Tabelle 23.3). Eine Erhöhung des Schwellenwertes von 3.0 (Standard-Wert für inSARa-basierte Analyse) auf 6.5 führt zu einer drastischen Reduktion der Kantenzahl im Gesamt-MCS-Schwellenwert-Netzwerk. Die Zahl der Kanten entspricht bei diesem Schwellenwert in etwa der Kantenzahl im inSARa-Schwellenwert-Netzwerk. Vergleicht man die Zahl der gemeinsamen Kanten, so haben das inSARa-basierte Schwellenwert-Netzwerk (Schwellenwert: 3.0) und das Gesamt-MCS-basierte Netzwerk 74,6% aller Kanten gemeinsam.

Tabelle 23.3. Abhängigkeit der Kantenzahl im Gesamt-MCS-Schwellenwert-Netzwerk und Anzahl der gemeinsamen Kanten mit dem inSARa-basierten Netzwerk vom Ähnlichkeits-Schwellenwert.

Ähnlichkeits-Schwellenwert (inSARa-MCSs)	Ähnlichkeits-Schwellenwert (gesamte MCS-Menge)	Zahl der Kanten in inSARa-Schwellenwert-Netzwerk	Zahl der Kanten in Gesamt-MCS-Schwellenwert-Netzwerk	Gesamtzahl an Kanten in beiden Schwellenwert-Netzwerken	Zahl an gemeinsamen Kanten	gemeinsame Kanten in Schwellenwert-Netzwerken (%)
3,0	3,0	511	4349	4349	511	11,7
3,0	4,0	511	2135	2137	509	23,8
3,0	5,0	511	1090	1097	504	45,9
3,0	6,0	511	622	664	469	70,6
3,0	6,5	511	519	590	440	74,6
3,0	6,75	511	481	569	423	74,3
3,0	7,0	511	448	549	410	74,7
3,0	7,25	511	425	541	395	73,0
3,0	7,5	511	391	527	375	71,2
3,0	8,0	511	346	522	335	64,2
3,0	9,0	511	267	512	266	52,0
3,0	10,0	511	226	511	226	44,2
3,0	11,0	511	183	511	183	35,8
3,0	12,0	511	156	511	156	30,5

Wie in Abbildung 23.17 zu erkennen, erreicht der prozentuale Anteil an gemeinsamen Kanten bei einem Schwellenwert von 7.0 sein Maximum mit fast 75%. Danach ist aufgrund der Abnahme der Kantenzahl im Gesamt-MCS-Schwellenwert-Netzwerk wieder eine Abnahme zu beobachten. Für einen Schwellenwert von 3.0 bis 6.0 ist der Anteil gemeinsamer Kanten ebenfalls aufgrund der großen Kantenzahl im Gesamt-MCS-Netzwerk relativ gering. Es ist jedoch zu sehen, dass annähernd alle Kanten des inSARa-Netzwerkes auch im Gesamt-MCS-Netzwerk vorkommen.

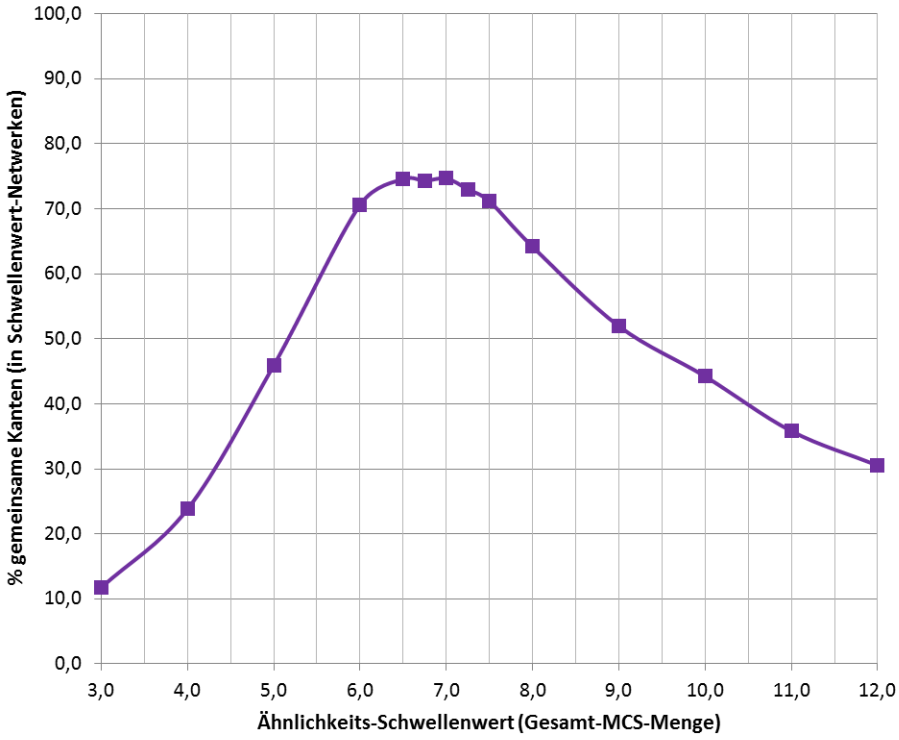


Abbildung 23.17. Vergleich: inSARa-MCSs und Gesamt-MCS-basiertes Target-Schwellenwert-Netzwerk. Gemeinsame Kanten in % der beiden Schwellenwert-Netzwerke in Abhängigkeit von dem Ähnlichkeits-Schwellenwert des Gesamt-MCS-Netzwerkes (Schwellenwert für inSARa-basiertes Netzwerk = 3.0).

Da die Ähnlichkeit für beide Analysen bei einem Schwellenwert von 7.0 am größten ist, wurde auf dieser Grundlage eine modifizierte Ähnlichkeitskarte (vgl. Abbildung 23.19) erstellt. Vergleicht man diese mit Abbildung 23.5 so sind deutliche Analogien im Farbmuster festzustellen.

In Abbildung 23.20 ist das hieraus resultierende Schwellenwert-Netzwerk dargestellt. Im Vergleich zu Abbildung 23.6 ist die Topologie leicht verändert. Die Gruppierung der Targetklassen ist ähnlich, die Verknüpfung der einzelnen Cluster ist etwas verändert und einige Verknüpfungen fehlen im Vergleich zum inSARa-Netzwerk (z.B. PDE und Adenosin-Rezeptor, COX2 und P38 oder Verknüpfung der Opioid-Rezeptoren mit dem Sigma-Rezeptor). Zudem sind einige zusätzliche Verknüpfungen zwischen den Targetklassen (v.a. über Orexin-1 und Orexin-2) feststellbar. Betrachtet man zusätzlich das Schwellenwert-Netzwerk, das bei einem Schwellenwert von 6.0 resultiert (vgl. Abbildung 23.21), so sind die meisten aus dem inSARa-basierten Schwellenwert-Netzwerk bekannten Ähnlichkeitsbeziehungen wieder zu finden. Hingegen sind noch eine Reihe weiterer Verknüpfungen vorhanden, die im inSARa-Netzwerk fehlen. Hierdurch resultiert ein sehr komplexes Netzwerk, das deutlich weniger unverknüpfte Knoten aufweist. Zudem zeichnet sich die Clusterbildung weniger deutlich ab, d.h. es sind im Vergleich zum inSARa-basierten Netzwerk mehr Intertargetklassen-Verknüpfungen vorzufinden. Da bei umso geringerem Schwellenwert die Wahrscheinlichkeit für falschpositive Kreuzreaktivitäten bzw. unbedeutsame Ähnlichkeitsbeziehungen steigt, müssten hier im Einzelfall die gemeinsamen MCSs und Liganden zur Beurteilung näher betrachtet werden.

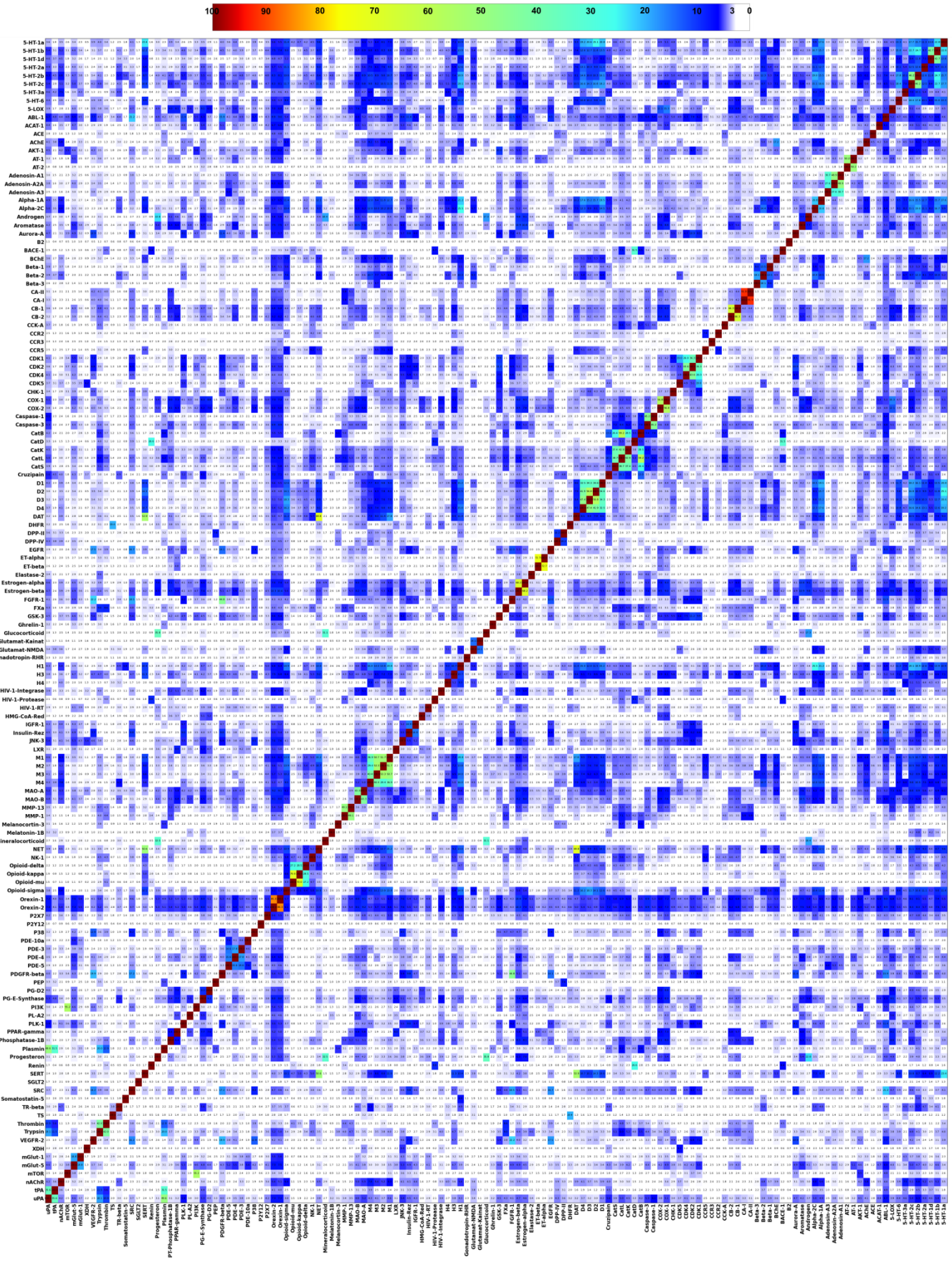


Abbildung 23.18. Ähnlichkeitskarte der gesamten MCS-Mengen der 140 BindingDB-Targets (gewichteter inSARa-TSim). weiß = geringe paarweise Ähnlichkeit, Ähnlichkeits-Zunahme von blau (ab 3.0) nach rot (vgl. Legende).

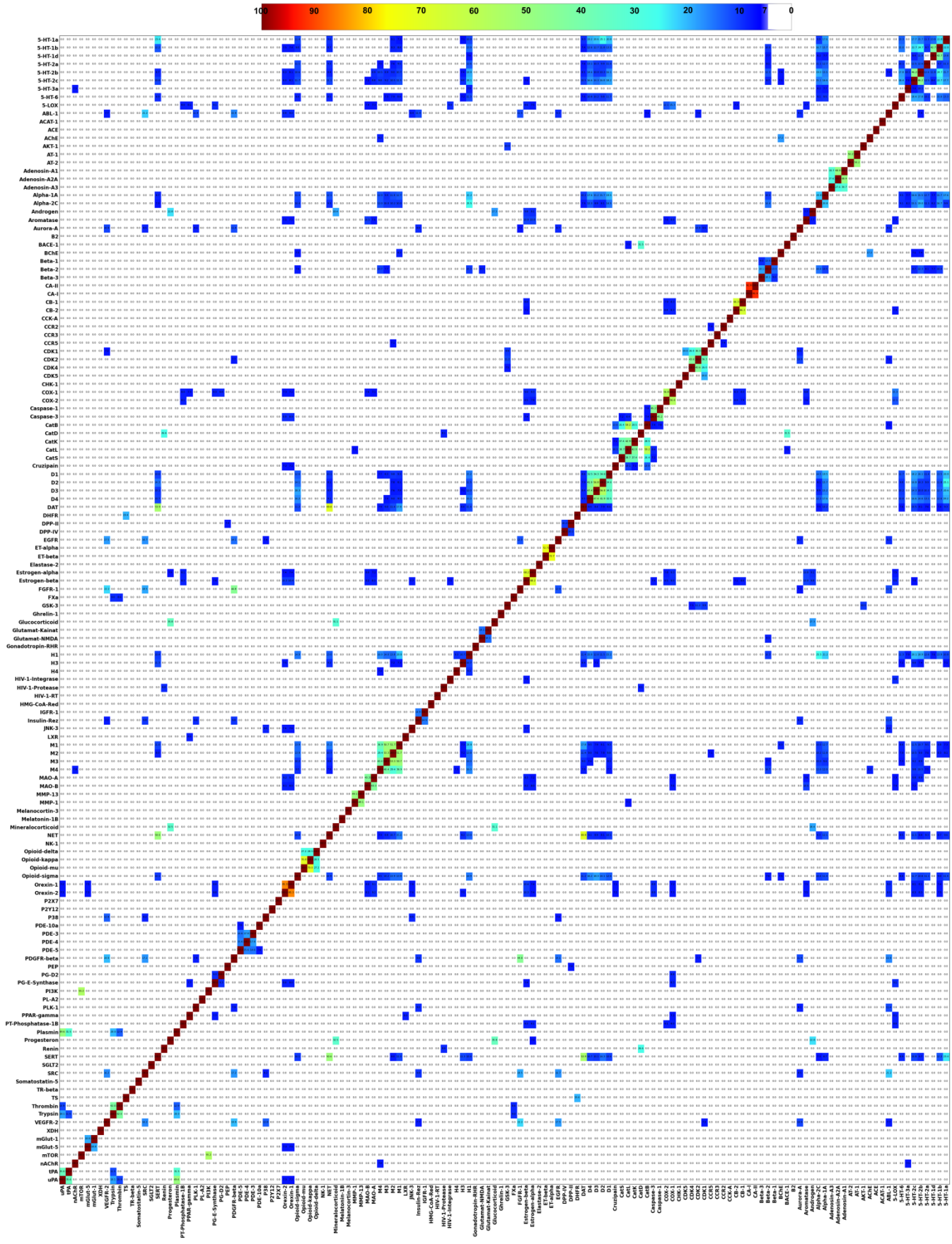


Abbildung 23.19. Ähnlichkeitskarte der gesamten MCS-Mengen der 140 BindingDB-Targets (gewichteter inSARa-TSim). weiß = geringe paarweise Ähnlichkeit, Ähnlichkeits-Zunahme von blau (ab 7.0) nach rot (vgl. Legende).

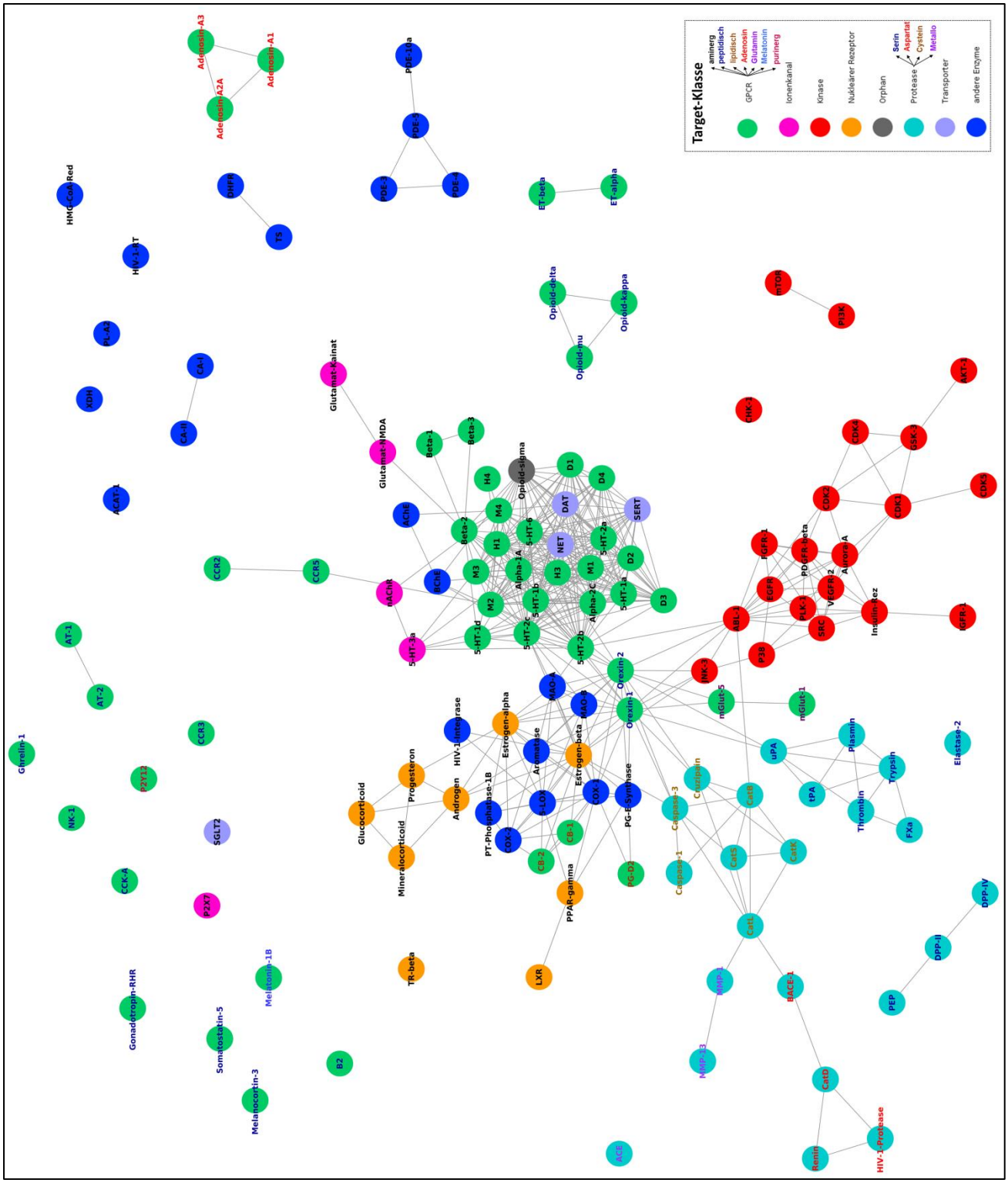


Abbildung 23.20. Schwellenwert-Netzwerk, das aus dem Ähnlichkeitsvergleich der gesamten MCS-Mengen der 140 BindingDB-Targets resultiert (Schwellenwert = 7.0, gewichteter inSARa-TSim). Knoten-Farbe gemäß Target-Klasse, Label-Farbe gemäß Subklasse, vgl. Legende).

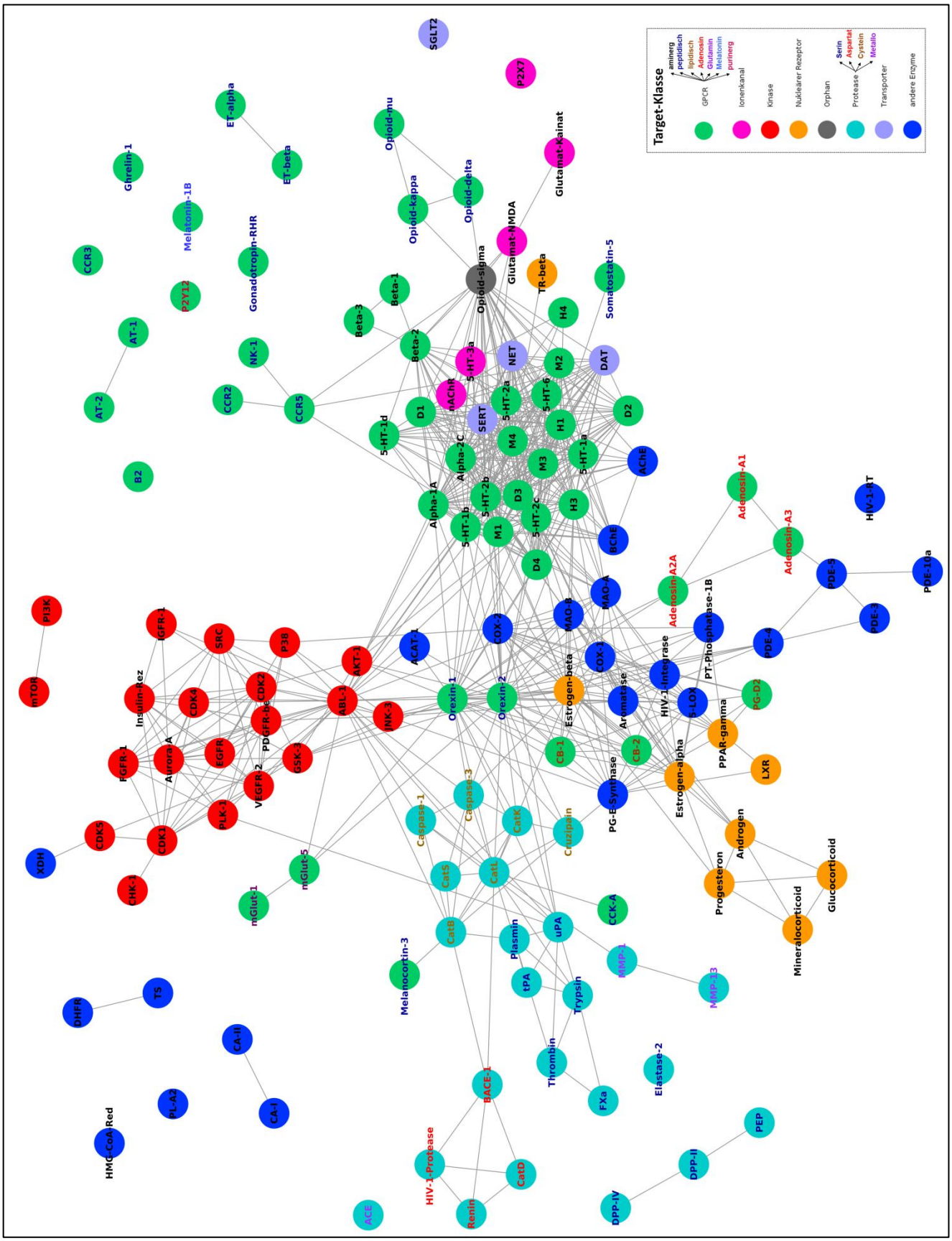


Abbildung 23.21. Schwellenwert-Netzwerk, das aus dem Ähnlichkeitsvergleich der gesamten MCS-Mengen der 140 BindingDB-Targets resultiert (Schwellenwert = 6.0, gewichteter inSARa-TSim). Knoten-Farbe gemäß Target-Klasse, Label-Farbe gemäß Subklasse, vgl. Legende).

Zusammenfassend lässt sich feststellen, dass die inSARa-basierten Ergebnisse den Ergebnissen, die aus der Analyse der gesamten MCS-Menge resultieren, ähnlich sind. Die in den Netzwerken repräsentierten MCSs sind somit in der Mehrzahl der Fälle repräsentativ für die Gesamtmenge bzw. die inSARa-Netzwerke enthalten die wichtigste MCS-Information. Je nach Wahl des Schwellenwertes zeigen sich Unterschiede bezüglich der gefundenen Ähnlichkeitsbeziehungen. Die inSARa-basierten Ergebnisse sind dadurch gekennzeichnet, dass etwas weniger komplexe Ähnlichkeitsbeziehungen resultieren, d.h. es finden sich v.a. weniger Verknüpfungen zwischen den verschiedenen Target-Klassen. Bei Berücksichtigung der gesamten MCS-Menge werden alle MCSs, die aus dem paarweisen Molekülvergleich resultieren berücksichtigt. Bei den inSARa-MCSs handelt es sich v.a. um MCSs, die die Mehrheit der Moleküle im Datensatz repräsentieren. Somit ist es möglich, dass einzelne Kreuzreaktivitäten unberücksichtigt bleiben. Jedoch ist bei Verwendung der gesamten MCS-Menge auch potentiell ein höheres Risiko für falschpositive Beziehungen gegeben (z.B. durch Ähnlichkeit basierend auf vielen kleinen MCSs, die aufgrund der Wurzel-Knoten-Auswahl im inSARa-Netzwerk nicht berücksichtigt wurden).

Der Vorteil der inSARa-basierten Variante ist, dass auf Basis der inSARa-Netzwerke, die für die Ähnlichkeit verantwortlichen MCS-Knoten inklusive zugehöriger Moleküle und benachbarter MCSs inklusive zugehöriger Moleküle sofort betrachtet werden können. Bei der Variante auf Basis der gesamten MCS-Menge ist der Vorteil, dass die z.T. etwas rechenaufwändigere Erstellung der inSARa-Netzwerke entfällt. Hingegen ist jedoch die Berechnung der Ähnlichkeitsbeziehungen deutlich rechenaufwändiger (v.a. bei Berücksichtigung von Substruktur-Beziehungen). Bei dieser Variante könnte man im Anschluss an die Analyse sich ebenfalls die gemeinsamen MCSs aus der Gesamtmenge extrahieren, mittels Sub- und Superstruktursuchen Beziehungen zwischen den gemeinsamen MCSs herstellen und mittels Substruktursuchen zugehörige Moleküle zur Beurteilung und Analyse der Ähnlichkeits-Beziehungen bzw. Kreuzreaktivitäten identifizieren.

Abschließend ist festzustellen, dass ligandbasierte Target-Analysen auf Basis von RG-MCSs ebenfalls möglich sind und in der Mehrzahl an Fällen sinnvolle oder experimentell bestätigte Beziehungen resultieren. Durch die andersartige Molekülrepräsentation (pharmakophore Eigenschaften) und die andersartige Ähnlichkeitserfassung (gemeinsame Substruktur statt Tc-Ähnlichkeit) hat der gezeigte Ansatz hat das Potential verfügbare Chemogenomik-Methoden zu ergänzen.

23.2. Diskussion

Ähnlichkeitsbeziehungen/Kreuzreaktivitäten

Die Analysen haben gezeigt, dass die Ähnlichkeitsbeziehungen im Schwellenwert-Netzwerk aus chemischer und pharmakologischer Sicht sinnvoll sind (vgl. Abschnitt 23.1.3). Die gefundene Kreuzreaktivitäten können in der Mehrheit (durch veröffentlichte experimentelle Testungen) bestätigt oder erklärt werden (vgl. Abschnitt 23.1.4: Ähnliche oder gleiche endogene Liganden, Multi-Target-Arzneistoffe, Entwicklungsgeschichte von Arzneistoff-Klassen). In einigen Fällen wären experimentelle Testungen zur Überprüfung der klinischen Relevanz der gefundenen Ähnlichkeit notwendig. Der Anteil an falsch-positiven Kreuzreaktivitäten oder Ähnlichkeitsbeziehungen ist relativ gering. Bei SEA-basierten Off-Target-Vorhersagen konnten beispielsweise fast 50% der Vorhersagen nicht experimentell bestätigt werden.[348] Bestätigende experimentelle Testungen werden daher niemals durch diese in-silico-Vorhersagen ersetzt werden können. Auch ist zu beachten, dass mit diesem Ansatz in der Regel nur Kreuzreaktivitäten erkannt werden, die auf einer deutlichen Überlappung von pharmakophoren Eigenschaften beruhen. Das sich ergebende Target-Netzwerk kann keine vollständige Darstellung aller existierenden pharmakologischen Beziehungen liefern. Aus einer fehlenden Beziehung zwischen zwei Targets im Netzwerk kann nicht darauf geschlossen werden, dass einzelne Liganden nicht doch duale Aktivität bzw. entsprechende Off-Target-Aktivitäten zeigen könnten. Eine hohe Zahl an falsch-negativen Beziehungen ist anzunehmen, denn Off-Target-Aktivitäten einzelner Moleküle werden bei diesem Ansatz nicht erfasst. Auch werden strukturelle Singletons, für die kein gemeinsamer MCS mit anderen Liganden desselben Targets existiert, bei diesem Ansatz im Gegensatz zu SEA nicht bei der Analyse von Target-Ähnlichkeiten berücksichtigt. Bei zukünftigen Weiterentwicklungen könnte man dieses Problem jedoch lösen, indem man für Singletons zusätzlich einen Vergleich mit den MCSs der anderen Targets einführen würde.

Vergleich des Prinzips mit anderen ligandbasierten Analysen

Im Gegensatz zu SEA oder anderen FP-basierten Ansätzen weist der RG-MCS-basierte Ansatz deutliche Vorteile auf. So ist wie in Abschnitt 23.1.5 beispielhaft aufgezeigt die Ähnlichkeit leicht auf Basis der Menge an gemeinsamen RG-MCSs bzw. der entsprechenden inSARa-Netzwerke leicht interpretierbar. Die Analyse von Gemeinsamkeiten ist auf diese Weise sehr intuitiv. Unspezifische oder unbedeutende Ähnlichkeiten lassen sich leicht erkennen. Substruktur-basierte Analysen sind bisher nur auf Basis der Molekülstruktur veröffentlicht. Dabei ist es häufig ein Problem, dass nur sehr kleine zusammenhängende gemeinsame Fragmente gefunden werden können und so Ähnlichkeiten u.U. nicht erkannt werden.[384] Auf Basis von RGs sind bisher keine Ansätze beschrieben. Da der Pharmakophor (wie in Kapitel 2.2.1 beschrieben) entscheidend für die biologische Aktivität von Molekülen ist, stellt dieses RG-MCS-basierte Verfahren einen sehr vielversprechenden Ansatz für Chemogenomik-Analysen dar. Die gefundenen sinnvollen Beziehungen bestätigen dies. Bei der RG-basierten Analyse ist aufgrund der Abstraktion von der Molekülstruktur zu erwarten, dass u.U. Ähnlichkeiten erkannt werden, die bei exaktem Molekülstruktur-Vergleich nicht deutlich werden. Wie bei den SAR-Analysen gilt auch hier, dass bei zu hohem Abstraktionslevel die Wahrscheinlichkeit für wenig sinnvolle Ähnlichkeiten zunimmt. Ein Problem dieses Ansatzes ist die starke Abhängigkeit der gefundenen

Ähnlichkeitsbeziehungen von der Datensatzgröße. Im Gegensatz zu SEA, wo ebenfalls eine starke Abhängigkeit von der Datensatzgröße besteht[345], ist hierbei eine zusätzliche Schwierigkeit, dass Datensätze gleicher Anzahl an Liganden unterschiedliche einzigartige MCS-Mengen mit zusätzlich unterschiedlichen MCS-Größen-Verteilungen ergeben können (vgl. Abschnitt 23.1.1). Diese Abhängigkeit ist bei der Interpretation von Ähnlichkeitsbeziehungen immer zu berücksichtigen. Absolute Ähnlichkeitswerte sind nur von beschränkter Aussagekraft. Bei Weiterentwicklung des Ansatzes müsste geprüft werden, ob sich die Datensatzgrößen-Abhängigkeit noch besser bei der Ähnlichkeitsberechnung durch Korrekturfaktoren ausgleichen oder in der Analyse berücksichtigen lässt. Die Berücksichtigung von Substruktur-Beziehungen stellt (wie in Abschnitt 23.1.1 untersucht) einen ersten vielversprechenden Ansatz zur partiellen Lösung dieses Problems dar.

Limitierung der Analyse durch die verfügbaren Daten

Wie auch bei der SAR-Analyse gilt auch bei Chemogenomik-Analysen, dass nicht nur die Methode, sondern auch die zugrunde liegenden Daten ein wichtiger Faktor für den Erfolg einer Analyse sind („Müll rein, Müll raus“-Prinzip[96]). Im Folgenden sollen einige Grenzen der durchgeführten Analyse, die auf die Datengrundlage zurückzuführen sind, aufgezeigt und diskutiert werden.

KALLIOKOSKI et al.[96] bemängelten, dass die Target-Annotationen von öffentlich zugänglichen Chemogenomik-Daten ungenügend sind. Dies konnte auch bei der in dieser Arbeit durchgeführten Analyse auf Basis der BindingDB bestätigt werden. Ein Problem ist vor allem die nicht-einheitliche Benennung von Targets. In der BindingDB konnten für viele Zielstrukturen Datensätze (auch mit unterschiedlicher Größe) unter unterschiedlichen Namen gefunden werden. Insbesondere bei den GPCRs hat sich eine große Variabilität bei der Benennung gezeigt (z.B. serotonin 2b receptor, 5-HT2b receptor, 5-hydroxytryptamine receptor 2B oder HTR2B (Genname)). Dies war auch ein Grund, warum die Datensätze nicht automatisch z.B. mittels Oracle-Datenbank SQL-Abfragen zusammengestellt wurden, sondern manuell ausgewählt wurden. Ein Vorteil der BindingDB ist jedoch, dass sie für jedes Molekül-Target-Paar die zugehörige speziesabhängige UniProtKB Accession Number mitliefert, sodass eine eindeutige Target-Identifizierung in der größten bioinformatischen Datenbank für Proteine UniProt (Universal Protein Database[545]) möglich ist. Die Verwendung einer standardisierten Nomenklatur (z.B. nach IUBMB Enzyme Nomenklatur[419], IUPHAR Nomenklatur für GPCRs[546–547] oder der speziesunabhängige UniProt Recommended Protein Name[548]) für die Benennung der verschiedenen Targets sollte zukünftig in den Datenbanken angestrebt werden. Unsicherheit bezüglich der Subtyp-Annotation von Targets sollte bei der Auswertung von Chemogenomik-Analysen berücksichtigt werden. KALLIOKOSKI et al. verweisen auf Fälle in ChEMBL, wo Moleküle an verschiedenen Subtypen als aktiv gelistet sind, obwohl die Bioaktivität an einer undefinierter Mischung an Subtypen bestimmt wurde.[96] Da viele Daten der BindingDB aus ChEMBL stammen, sind ähnliche Fehler bei dieser Analyse ebenfalls nicht auszuschließen. Des Weiteren enthalten die verwendeten Datensätze nicht nur Daten aus verschiedenen Assays, sondern auch zum Teil von verschiedenen Spezies. Die COX-2-Daten beispielsweise stammen von 3 verschiedenen Spezies (Mensch, Hausschaf, Hausmaus). Da bei einigen Targets speziesabhängige Unterschiede in der Bioaktivität beschrieben sind[307, 549], kann hier ebenfalls ein systematischer Fehler resultieren.

Des Weiteren sei darauf verwiesen, dass für diese Analysen Daten aus Bindungs-Assays verwendet werden. Wie in Kapitel 2.2.2 beschrieben, wird hierbei nur die Affinität zum Target nicht jedoch die Art und Weise der Wirkung bestimmt. D.h. in den Datensätzen befinden sich z.T. sowohl Agonisten als auch Antagonisten für ein Target (z.B. Cannabinoid-Rezeptor). Das bedeutet, dass die pharmakologische Auswirkung einer gefundenen Kreuzreaktivität in einigen Fällen schwer vorhersagbar ist. Auch kann es sein, dass eine Kreuzreaktivität u.U. nicht von praktischer Relevanz ist, weil Bindungsaffinität und pharmakologischer Effekt nicht korrelieren (vgl. Beispiel in Kapitel 2.2.2). Um die genannten Probleme zu lösen, wären zusätzliche Annotationen der Moleküle anhand von Daten aus funktionellen Assays hilfreich, sodass z.B. getrennte Datensätze für beispielsweise Agonisten und Antagonisten erstellt werden können. In den Bioaktivitäts-Datenbanken sind zurzeit solche Informationen jedoch bisher nicht (BindingDB) oder nur in Einzelfällen (ChEMBL) hinterlegt.

Die Ergebnisse solcher Analysen sind stark von den Liganden in den verwendeten Datensätzen abhängig. Je mehr Daten zur Verfügung stehen bzw. in die Analyse einbezogen werden, desto besser. Bei dieser Analyse wurden für ein Target entweder nur K_i oder IC_{50}-Daten verwendet, da die verwendeten Datensätze primär für die Anwendung in der SAR-Analysen zusammengestellt worden sind. Für SAR-Analysen ist eine Vermischung verschiedener Bioaktivitätsdaten-Typen nicht sinnvoll. Ist man jedoch nur an einer Chemogenomik-Analyse interessiert, ist ein Mischen von K_i/IC_{50}-Daten möglich (vgl. KALLIOSKI et al.[99]). Für zukünftige Arbeiten wäre dies eine Möglichkeit die Datengrundlage zu vergrößern und die Analyse noch robuster zu machen bzw. noch mehr Targets in die Analyse einzubeziehen.

Die optimale Ausgangsbedingung für solche Analysen wären an allen Targets experimentell getestete Liganden. Da experimentelle Testungen jedoch sehr teuer sind, sind solche Voraussetzungen für Analysen in diesem großen Maßstab auch zukünftig (zumindest im öffentlichen Bereich) nicht zu erwarten. Dieser Punkt wird noch lange Zeit die „Achillesferse“[330] entsprechender Analysen bleiben.

23.3. Zusammenfassung

Die Ergebnisse dieser Analyse zeigen, dass die Codierung der Moleküle mittels RG bedeutsam zu sein scheint und auch die Ähnlichkeitserfassung über die Bestimmung des MCS sinnvoll ist. Ähnlichkeiten zwischen Targets ähnlicher (Sub-)Familien sind erkennbar. Kreuzreaktivitäten lassen sich zumeist gut erklären bzw. durch Literaturrecherchen bestätigen.

Die guten Übereinstimmungen der Ergebnisse beim Vergleich der inSARa-basierten mit der Gesamt-MCS-basierten Analyse zeigen, dass die wichtigste Information der Datensätze ebenfalls in den inSARa-Netzwerken repräsentiert wird.

Es konnte gezeigt werden, dass Vergleiche auf Grundlage des MCS eine intuitive Alternative der Analyse von Beziehungen zwischen Zielstrukturen darstellen. Der Vorteil der direkten Interpretierbarkeit ist dabei sehr entscheidend. Ein Problem, das bei zukünftigen Weiterentwicklungen zu beheben ist, stellt die starke Datensatzgrößen-Abhängigkeit des Verfahrens dar. Für absolute Ähnlichkeitsanalysen sollte hier noch ein weiterer

Korrekturfaktor (zusätzlich zur Berücksichtigung von Substruktur-Beziehungen) entwickelt werden.

RGs haben den zusätzlichen Vorteil der Fokussierung auf pharmakophore Eigenschaften, die potentiell an der Ligand-Rezeptor-Interaktion beteiligt sind. Ähnliche Targets sollten somit auch ähnliche Bindetascheneigenschaften aufweisen. Der Ansatz könnte in der prospektiven Anwendung somit insbesondere für Targets mit bisher nicht-aufgeklärter Struktur aber bekannten Liganden interessant sein und beim Auffinden neuer Liganden oder der Ableitung von Target-Eigenschaften helfen.

Diese Analyse kann wie beispielhaft aufgezeigt zudem bei der Erklärung von bekannten UAWs oder pharmakologischen Effekten (durch Identifizierung von unbekannten Off-Targets) hilfreich sein. Bisher wurden bei diesem Ansatz ganze Datensätze verglichen. Eine Fokussierung auf einzelne Moleküle im Hinblick auf Vorhersagen von potentiellen (On/Off-) Targets wäre als Weiterentwicklung ebenfalls denkbar (vgl. Ausblick in Kapitel 25).

Zusammenfassend lässt sich nicht nur feststellen, dass die für die SAR-Interpretation entwickelten inSARa-Netzwerke ebenfalls wertvolle Information im Hinblick auf Polypharmakologie enthalten, sondern auch dass dieser neuartige RG-MCS Ansatz das Potential für die Ergänzung verfügbarer chemogenomischer Methoden aufweist.

24. Zusammenfassung

Aufgrund des fortschreitenden technologischen Fortschritts ist eine fortlaufend steigende Menge an Bioaktivitätsdaten verfügbar. Da hier wichtige Information enthalten sein kann, die die Entwicklung neuer Arzneistoffe in verschiedenen Phasen entscheidend beschleunigen kann, ist die systematische, automatisierte Auswertung und kompakte Visualisierung dieser großen Datenmengen eine zentrale Herausforderung der heutigen Arzneistoffentwicklung, die es mit geeigneten computergestützten Methoden zu lösen gilt.

Die Hauptmerkmale, die die in dieser Arbeit entwickelte inSARa-Methode (intutive networks for Structure-Activity Relationships analysis) von bisherigen Ansätzen zur Analyse und Visualisierung von Struktur-Aktivitäts-Beziehungen (SAR) unterscheiden, sind hierarchische Netzwerke klar-definierter Substruktur-Beziehungen auf Basis von gemeinsamen pharmakophoren Eigenschaften. Durch Kombination des Konzeptes des „reduzierten Graphen“ (RG) mit dem intuitiven Konzept der „maximal gemeinsamen Substruktur“ (MCS) resultiert ein besonderer Synergismus für die SAR-Interpretation. Dieser ermöglicht, dass der medizinische Chemiker leicht gemeinsame bzw. bioaktivitätsbeeinflussende molekulare (pharmakophore) Merkmale in großen, auch strukturell diverseren Datensätzen, die aus Hunderten oder Tausenden von Molekülen bestehen, erfassen kann.

Beim Analysieren von Datensätzen aktiver Moleküle einzelner Zielstrukturen haben sich die ohne Berücksichtigung von Bioaktivitätsinformation aufgebauten inSARa-Netzwerke als wertvoll für verschiedene essentielle Aufgaben der SAR-Analyse erwiesen. Neben gemeinsamen pharmakophoren Eigenschaften lassen sich so auf Grundlage einfacher Regeln bioisosterer Austausch, sprunghafte SARs oder „SAR Hotspots“ und sogenannte „Activity Switches“, wo die Variation bestimmter pharmakophorer Eigenschaften eine systematische Veränderung der Bioaktivität bewirkt, erkennen. Sowohl mittels interaktiver inSARa-Netzwerk-Navigation als auch durch automatisierte Analyse (inSARa^{auto}) können die verschiedenen Typen an SAR-Information identifiziert werden.

In dieser Arbeit konnten verschiedene Parameter zur Optimierung der Netzwerke aufgezeigt werden. Ebenfalls konnte in verschiedenen vergleichenden Analysen eine Komplementarität oder Überlegenheit zum häufig verwendeten Ansatz der Fingerprint-basierten Ähnlichkeitsanalyse gezeigt werden. Der inSARa Hybrid Ansatz, der inSARa in verschiedenen Varianten mit Fingerprint-basierten Ähnlichkeits-Netzwerken kombiniert, zeigt zudem die Vorteile auf, die aus der Kombination beider Prinzipien resultieren können. Analog zum Fingerprint-basierten SAR-Index ermöglicht der auf inSARa^{auto} aufbauende SARdisco Score die globale Charakterisierung der Verteilung von SAR-(Dis-)Kontinuität in inSARa-Netzwerken verschiedener Datensätze.

Polypharmakologie hat eine vielseitige Bedeutung für die Entwicklung neuer Arzneistoffe. Der Vergleich der inSARa-Netzwerke verschiedener Zielstrukturen auf Basis der Schnittmenge an RG-MCSs hat gezeigt, dass dieses Konzept aufgrund seiner einfachen Interpretierbarkeit und Fokussierung auf Eigenschaften, die in die Protein-Ligand-Bindung involviert sind, eine vielversprechende Ergänzung verfügbarer Chemogenomik-Ansätze zur ligandbasierten Analyse von Target-Ähnlichkeiten und zur Identifizierung von Kreuzreaktivitäten darstellt. Potentielle Off-Target-Beziehungen und (un)erwünschte Arzneimittelwirkungen können so vorhergesagt oder erklärt werden. Zudem lassen sich selektivitätsentscheidende Merkmale oder Eigenschaften ableiten, die die rationale

Entwicklung von Multi-Target-Arzneistoffen zur Therapie komplexer Erkrankungen ermöglichen. Zudem lassen sich so potentiell für Targets, für die keine Strukturinformation verfügbar ist, ähnliche Targets und folglich potentielle neue Arzneistoffe oder Bindetaschen-Eigenschaften vorhersagen.

Zusammenfassend lässt sich feststellen, dass Zufall in der Vergangenheit immer ein wichtiger Faktor in der Arzneistoffentwicklung war und auch in Zukunft bleiben wird. Durch Kombination verfügbarer Ansätze zur SAR-Analyse und Chemogenomik-Analyse kann wichtiges Wissen gewonnen werden, das den schwierigen Prozess der Arzneistoffentwicklung unterstützen und weiter rationalisieren kann. Durch die vielseitige Anwendbarkeit und die zugrundeliegende intuitive, komplementäre Methodik kann somit von dem in dieser Arbeit entwickelten inSARa-Ansatz ein wichtiger Beitrag zur Entwicklung neuer und sicherer Arzneistoffe erwartet werden.

25. Ausblick

In dieser Arbeit wurden die grundlegenden Prinzipien für inSARa erarbeitet. Anhand retrospektiver Analysen auf Basis einer begrenzten Zahl frei verfügbarer Daten wurde inSARa im möglichen Rahmen getestet und optimiert. Hierbei konnte gezeigt werden, dass das der Methode zugrunde liegende Konzept Potential sowohl für die Anwendung im Bereich der SAR- und als auch im Bereich der Chemogenomik-Analyse aufweist. Nichtsdestotrotz bleiben noch viele Ansatzpunkte und Fragestellungen für weiterführende Arbeiten, die im Folgenden kurz aufgezeigt werden sollen.

Da inSARa bisher nur im begrenzten Umfang getestet werden konnte, wäre ein nächster wichtiger Schritt die Anwendung in großem Umfang durch verschiedene Benutzer auf unterschiedlichsten Daten. Entsprechende Rückmeldungen von Anwendern könnten zur weiteren Optimierung und Erweiterung der Methode bzw. Anpassung der Methode an die spezifischen Bedürfnisse der entsprechenden Benutzer oder unberücksichtigten Fallbeispielen (z.B. nicht-berücksichtigte chemische Funktionalitäten bei der RG-Erzeugung) im Quellcode beitragen.

Bisher ist inSARa aufgrund der Nutzung der kommerziellen OEChem TK Programmierbibliothek für die Implementierung nicht frei verfügbar. Eine Reimplementierung von inSARa als Open-Source Variante ist jedoch auf Grundlage des beschriebenen Netzwerk-Erzeugungs-Algorithmus mit entsprechenden freiverfügbaren Bibliotheken (z.B. RDKit[550]) möglich. Dies würde das Spektrum der potentiellen Anwender enorm erweitern und die Methode sowohl für die Anwendung in akademischen Gruppen als auch in der Pharmaindustrie attraktiv machen.

Das ultimative Ziel zur Beurteilung der tatsächlichen Leistungsfähigkeit der Methode wäre die prospektive Anwendung von inSARa für das Ableiten von SARs für die Entwicklung neuer Arzneistoffe. Die Aussagekraft der durchgeführten retrospektiven SAR-Analysen einer Vielzahl von öffentlich verfügbaren Datensätzen ist limitiert. Der reale Nutzen für die Arzneistoffentwicklung ist noch nachzuweisen.

Zudem wäre neben einer weiteren Optimierung der RG-Definition eine Untersuchung alternativer Möglichkeiten zur Reduktion der Netzwerk-Komplexität ein wichtiger Aspekt zur

Verbesserung der Interpretierbarkeit bzw. Leistungsfähigkeit der Methode. Weitere sinnvolle Projekte zur Verbesserung der Bedienung und Anwendbarkeit für den Benutzer wären das Programmieren eines Plug-In für Cytoscape zum vollautomatisierten Laden der Netzwerk-Dateien und zur Erzeugung des Layouts. Zur Steigerung der Intuitivität wäre eine weitere Optimierung der Visualisierung z.B. durch Übersetzen der aus Pseudoatomen bestehenden RGs in entsprechend codierte Eigenschaften ein weiterer interessanter Aspekt. Eine Erweiterung von inSARa^{auto} im Hinblick auf das automatisierte Hervorheben interessanter Netzwerk-Pfade (analog z.B. der „SAR Pathways“, vgl. Abschnitt 2.6.4) könnte die Netzwerk-Interpretation ebenfalls weiter verbessern.

Des Weiteren sind verschiedene algorithmische Abwandlungen oder Erweiterungen der Methode denkbar. So könnte inSARa z.B. wie in Kapitel 10.3 bereits angedeutet, dahingehend abgewandelt werden, dass bei der MCS-Bestimmung auch der nicht-zusammenhängende Fall berücksichtigt wird. Jedoch müsste dann die Netzwerk-Struktur entsprechend angepasst werden. Zudem werden die inSARa-Netzwerke bisher komplett unüberwacht ohne Ausnutzung der eigentlich verfügbaren Bioaktivitätsinformation erzeugt. Eine Abwandlung des Netzwerk-Algorithmus unter Berücksichtigung der vorhandenen Bioaktivitätsinformation wäre ähnlich den Prinzipien, die der Erzeugung von Entscheidungsbäumen zugrunde liegen, denkbar und ein weiterer interessanter Aspekt für zukünftige Weiterentwicklungen. Durch Einführung dieser Elemente des überwachten Lernens in den Netzwerk-Aufbau könnte ggf. das leichtere Ableiten genereller SAR-Regeln für einzelne Targets ermöglicht werden. Entsprechende Vorversuche haben jedoch gezeigt, dass der Erhalt der hierarchischen Netzwerkstruktur unter der Prämisse, möglichst viele Datensatzmoleküle in dem Netzwerk zu repräsentieren, nur schwer zu realisieren ist.

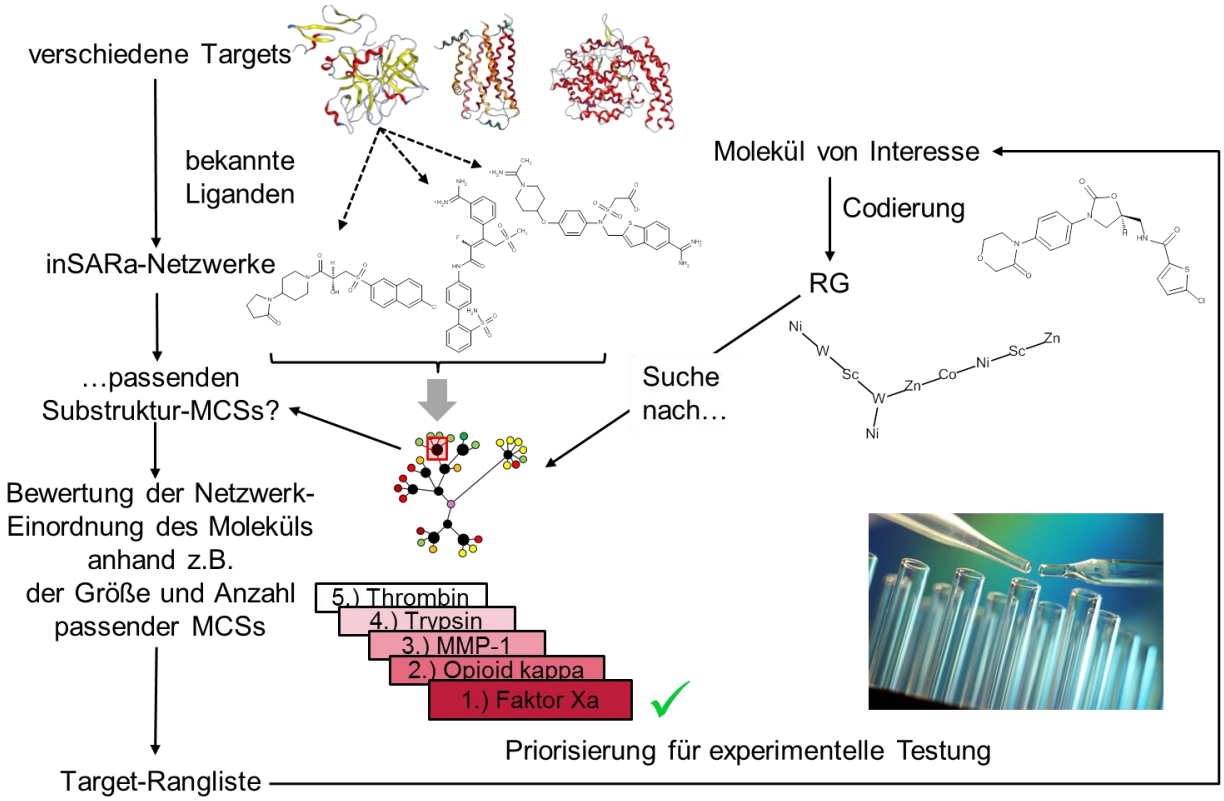


Abbildung 25.1. Ausblick auf weitere Anwendungsgebiete: Prinzip der Vorhersage von Targets auf Basis von inSARa-Netzwerken.

Ein interessanter Ansatzpunkt für weiterführende Arbeiten wäre zudem die Anwendung von inSARa im Bereich der Zielstruktur-Vorhersage („Target Fishing“, vgl. Abbildung 25.2 und Kapitel 8.2). Oftmals steht man vor dem Problem, dass man einen phänotypischen Effekt beobachten kann, es aber unbekannt ist, welche Zielstruktur moduliert wird. Dies ist z.B. im Bereich der Naturstoffe, einer vielversprechenden Quelle für neue Arzneistoffe, von großem Interesse. Aber auch im Bereich der Pflanzenschutzforschung, wo Moleküle z.B. in-vivo im Feldversuch einen fungiziden, herbiziden oder insektiziden Effekt zeigen, jedoch die genaue Zielstruktur zunächst unbekannt ist, ist dies von großem Interesse. Die Vorhersage des Targets oder der entsprechenden Target-Klasse (z.B. GPCR, Ionenkanal, Kinase) bzw. ein Ranking potenzieller Zielstrukturen könnte im Hinblick auf die Priorisierung von entsprechenden in-vitro Assays von hoher Relevanz sein und zum einen eine Kosten als auch eine deutliche Zeitersparnis für den kompletten Entwicklungsprozess einer Substanz darstellen. Da erste Vorversuche in diesem Bereich vielversprechend sind, ist eine entsprechende weiterführende Arbeit auf diesem Gebiet unter Berücksichtigung des praktischen Hintergrundes sehr wünschenswert. Abbildung 25.1 zeigt das Prinzip der Target-Vorhersage auf Basis von inSARa-Netzwerken auf. Beim Einsatz im Bereich der Naturstoffe ist (wie in Abschnitt 23.1.5 bereits andiskutiert) zu berücksichtigen, dass die RG-Definition an die molekularen Besonderheiten von Naturstoffen ggf. zuvor angepasst werden müsste.

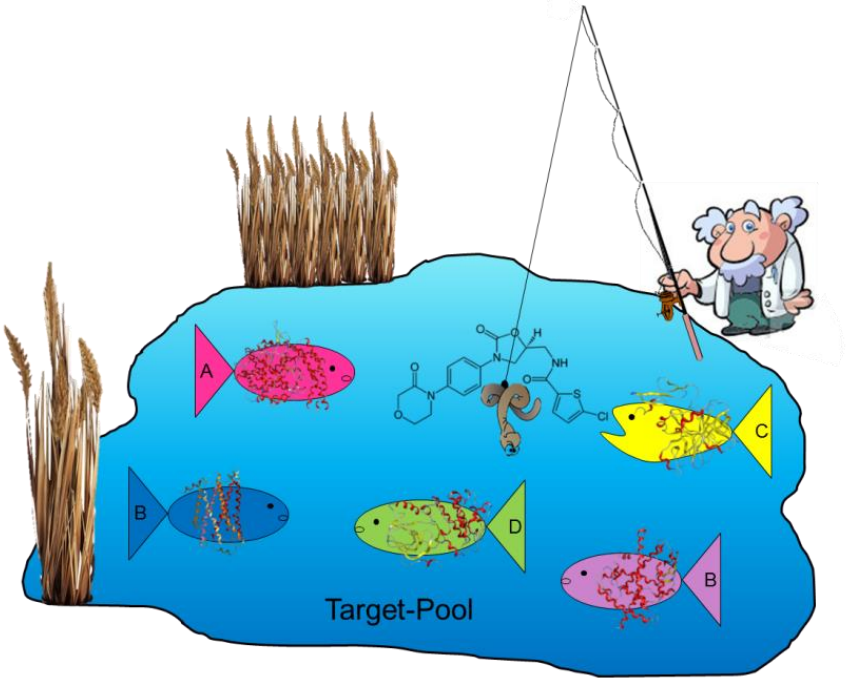


Abbildung 25.2. Prinzip des „Target Fishing“, d.h. der Vorhersage potentieller (Off-/On-)Targets für ein bestimmtes Molekül von Interesse.

IV. Anhang

26. Einstellungen und zusätzliche Abbildungen/Tabellen

26.1. Cytoscape: Layout-Einstellungen

Tabelle 26.1: Optimierte ‘Layout Settings’ in Cytoscape (engl.) für die Layout-Erstellung der inSARa-Netzwerke. Nach der Layout-Erzeugung ist eine Reskalierung des Netzwerkes auf ¼ empfehlenswert.

Layout Algorithm	Force-Directed Layout
Standard settings	default settings (Verbesserung des Layouts in einigen Fällen durch Weglassen des Hakens bei der Option “Partition graph before layout”, die in den Standard-Einstellungen gesetzt ist, danach Reskalierung meist nur auf ½ notwendig)
Edge Weight Settings	default settings
Algorithm settings	tuned
Default Spring Coefficient	9.99E-6 (default: 9.99E-5)
Default Spring Length	50.0 (default)
Default Node Mass	10.0 (default: 3.0)
Number of Iterations	10000 (default: 100)
Force deterministic layouts (slower)	yes (default: no)

26.2. Weitere Datensatz-Charakteristika

Tabelle 26.2. Übersicht über weitere Charakteristika der Datensätze aus der BindingDB. Abkürzung: Kont-norm = globaler Kontinuitäts-Score, Disk-norm = globaler Diskontinutitäts-Score

Abkürzung	Target	Target-Klasse	Anzahl an Molekülen nach Vorbereitung und RG-Erzeugung	Bioaktivitäts-Wert	Bioaktivitäts-Verteilung			Fingerprint-Ähnlichkeit Tc (MACCS Keys)			Fingerprint-Ähnlichkeit Tc (ECFP4)			Globaler SAR-Index (MACCS Keys 0.75) (berechnet mit SARANEA)			SAR-Typ (Klassifikation nach Kapitel 2.4.3)	SARdisco-Wert
					Min	Max	Mean	Min	Max	Mean	Min	Max	Mean	Kont-norm	Disk-norm	SARI		
FXA	Faktor Xa	Enzym (Protease)	1736	pIC_{50}	4,07	10,70	7,13	0,15	1,0	0,51	0,01	1,0	0,15	0,78	0,88	0,50	"heterogen-relaxed"	0,68
COX2	Cyclooxy-genase-2	Enzym	2349	pIC_{50}	4,01	11,22	6,33	0,0	1,0	0,41	0,0	1,0	0,14	0,95	0,77	0,60	"heterogen-relaxed"	0,67
CB1	Cannabinoid Rezeptor 1	GPCR	1957	pK_i	4,10	9,96	6,96	0,0	1,0	0,40	0,0	1,0	0,12	0,97	0,54	0,71	konti-nuierlich	0,70
CDK2	Cyclin-dependent Kinase 2	Enzym (Kinase)	1575	pIC_{50}	4,01	9,52	6,72	0,09	1,0	0,45	0,02	1,0	0,11	0,89	0,56	0,67	konti-nuierlich	0,77
P38	MAP Kinase p38 alpha	Enzym (Kinase)	2446	pIC_{50}	4,06	10,22	7,04	0,09	1,0	0,45	0,01	1,0	0,12	0,88	0,45	0,71	konti-nuierlich	0,79
THR	Thrombin	Enzym (Protease)	2852	pK_i	4,00	12,19	6,76	0,09	1,0	0,49	0,0	1,0	0,14	0,83	0,94	0,45	"heterogen--relaxed"	0,61

26.3. Analyse unspezifischer Ähnlichkeit in der ZINC Datenbank

Tabelle 26.3: Ausschlussliste. Liste aller RG MCSs, die eine Häufigkeit von mindestens 0.1 % in Zufalls-Molekülpaaren aufweisen. Die Häufigkeit wird berechnet aus der Anzahl des Auftretens geteilt durch 10^6 (= Anzahl an gezogenen Zufallspaaren).

Rang	Häufigkeit [%]	RG MCS SMILE
1	6.53	[Co]([Ni])[Zn]
2	6.4	[Sc][Ni][Zn]
3	4.56	[Co][Ni][Zn]
4	4.02	[Sc]([Ni])[Zn]
5	3.61	[Sc]([Zn])[Zn]
6	2.2	[Sc]([Ni][Zn])[Zn]
7	2.04	[Sc][Co][Ni]
8	1.29	[Co]([Ni][Zn])[Zn]
9	1.26	[Ni]([Zn])[Zn]
10	1.08	[Sc][Zn][Co][Ni]
11	1.02	[Sc][Zn][Ni]
12	1.02	[Sc][Ni][Co]
13	0.85	[Ni][Zn][Ni]
14	0.76	[Sc]([Co][Ni])[Zn]
15	0.73	[Zn][Nb][Zn]
16	0.7	[Sc][Co][Ni][Zn]
17	0.69	[Sc]([Ni])[Ni][Zn]
18	0.65	[V]([Ni])[Zn]
19	0.63	[V]([Zn])[Zn]
20	0.46	[Co][Ni][Zn][Ni]
21	0.45	[Sc][Ni][Co][Zn]
22	0.44	[V][Zn][Ni]
23	0.43	[Co]([Ni])[Zn][Ni]
24	0.41	[V][Ni][Zn]
25	0.4	[Sc]([Ni])[Ni]
26	0.4	[Sc]([Co][Ni])[Ni]
27	0.39	[Sc]([Ni][Zn])[Ni][Zn]
28	0.37	[Ni]([Zn])[Hf]
29	0.32	[Sc]([Ni][Co])[Zn]
30	0.31	[Sc][Zn][Nb]
31	0.3	[Sc]([Ni])([Zn])[Zn]
32	0.29	[Sc][Ni][Zn][Ni]
33	0.28	[Sc]([Co][Ni])[Ni][Zn]
34	0.24	[Sc]([Co][Ni][Zn])[Zn]
35	0.24	[Ni][Zn][Ni][Zn]
36	0.23	[Ni][Zn][Nb]
37	0.2	[Sc][V][Zn]
38	0.2	[Co][Ni][Co][Zn]

39	0.2	[Sc][Zn][Co][Ni][Zn]
40	0.19	[Sc][Zn][Nb][Zn]
41	0.19	[Co][Ni][Hf]
42	0.18	[Sc]([Zn])[Zn][Ni]
43	0.17	[Sc]([V])[Zn]
44	0.16	[Sc]([Zn])[Zn][Co][Ni]
45	0.16	[V][Zn][Co][Ni]
46	0.16	[V][Co][Ni]
47	0.14	[Sc][Zn][Ni][Zn]
48	0.14	[Sc][Zn][Ni][Co]
49	0.14	[Zn][Nb]([Zn])[Zn]
50	0.13	[Co]([Ni])[Zn][Nb]
51	0.13	[Sc]=[V][Zn]
52	0.13	[V]([Ni][Zn])[Zn]
53	0.13	[Sc]([Ni][Co][Zn])[Zn]
54	0.13	[Sc][Ni][Hf]
55	0.12	[Sc]([Co][Ni][Zn])[Ni]
56	0.12	[Sc]([Ni][Zn][Ni])[Zn]
57	0.11	[Ni][W][Zn]
58	0.11	[V][Ni][Co]
59	0.11	[Co]([Ni])[Hf]
60	0.1	[Sc]([Ni])([Ni])[Zn]

26.4. RG-Größen-Verteilung in den analysierten Datensätzen

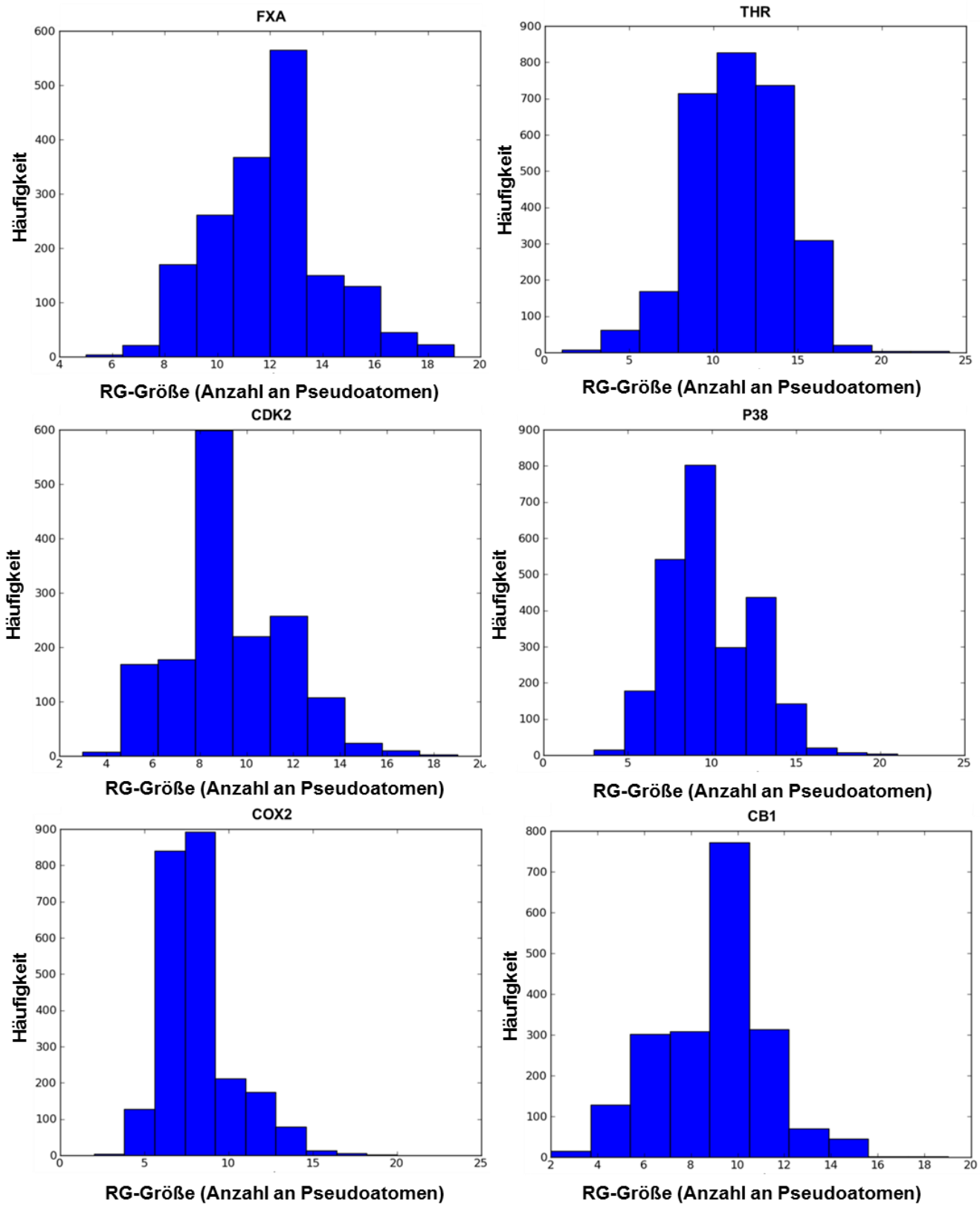


Abbildung 26.1. RG-Größen-Verteilung der analysierten Datensätze.

26.5. Vergleich verschiedener Ähnlichkeits-Koeffizienten für den ligandbasierten Target-Vergleich

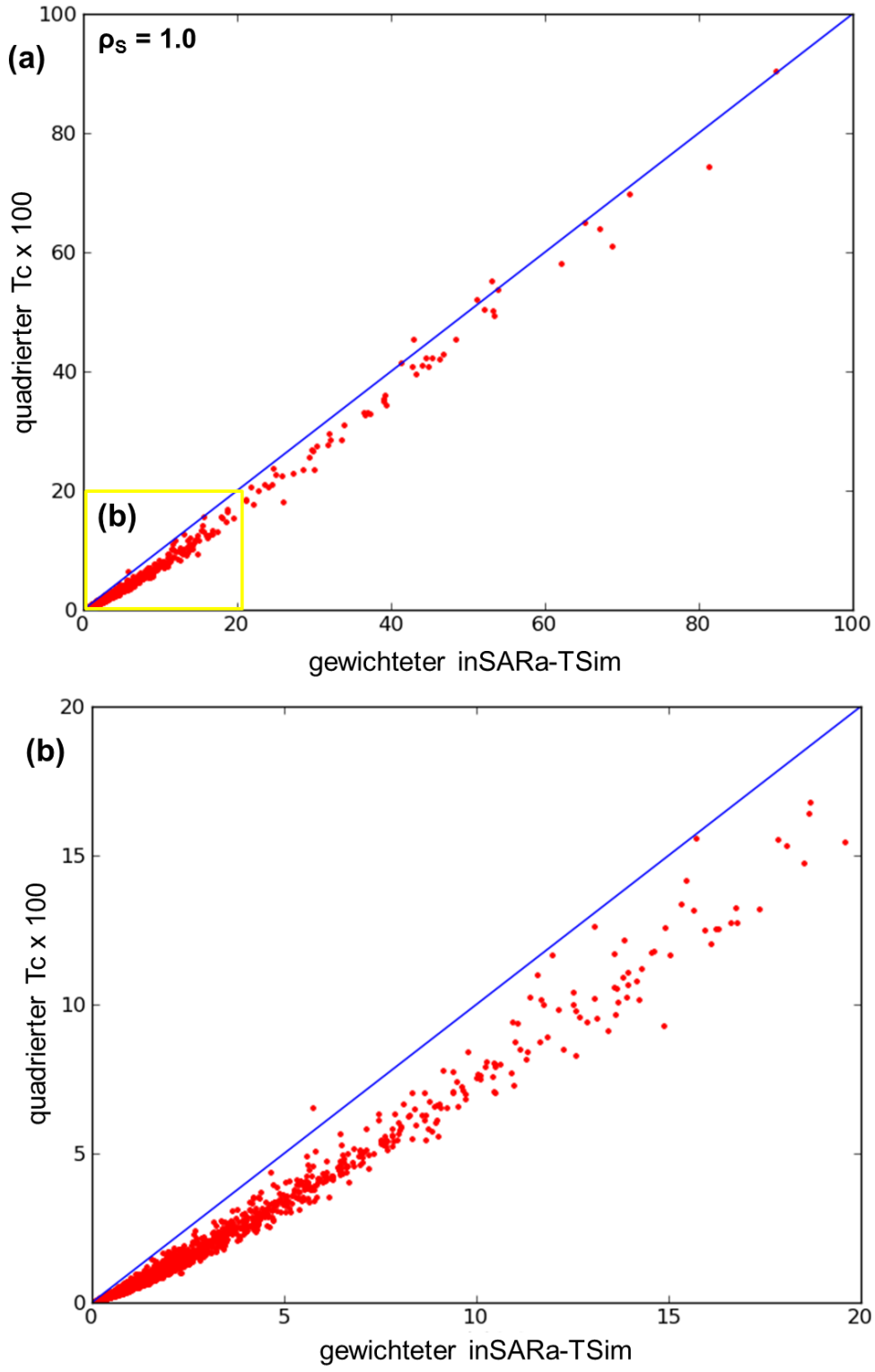


Abbildung 26.2. Korrelation (Spearman-Rang-Korrelationskoeffizient ρ_S) zwischen dem gewichteten inSARa-TSim und dem quadrierten Tc (auf den Bereich von 0 bis 100 skaliert) beim paarweisen Vergleich aller Targets. Eigenvergleiche (=100) sind nicht berücksichtigt. (b) stellt eine Vergrößerung des gelb markierten Bereiches aus (a) dar. Zum Vergleich: Pearson-Korrelationskoeffizient r_P = 0.99.

26.6. Ergebnisse des „Selbstähnlichkeitstestes“ (P38/COX2)

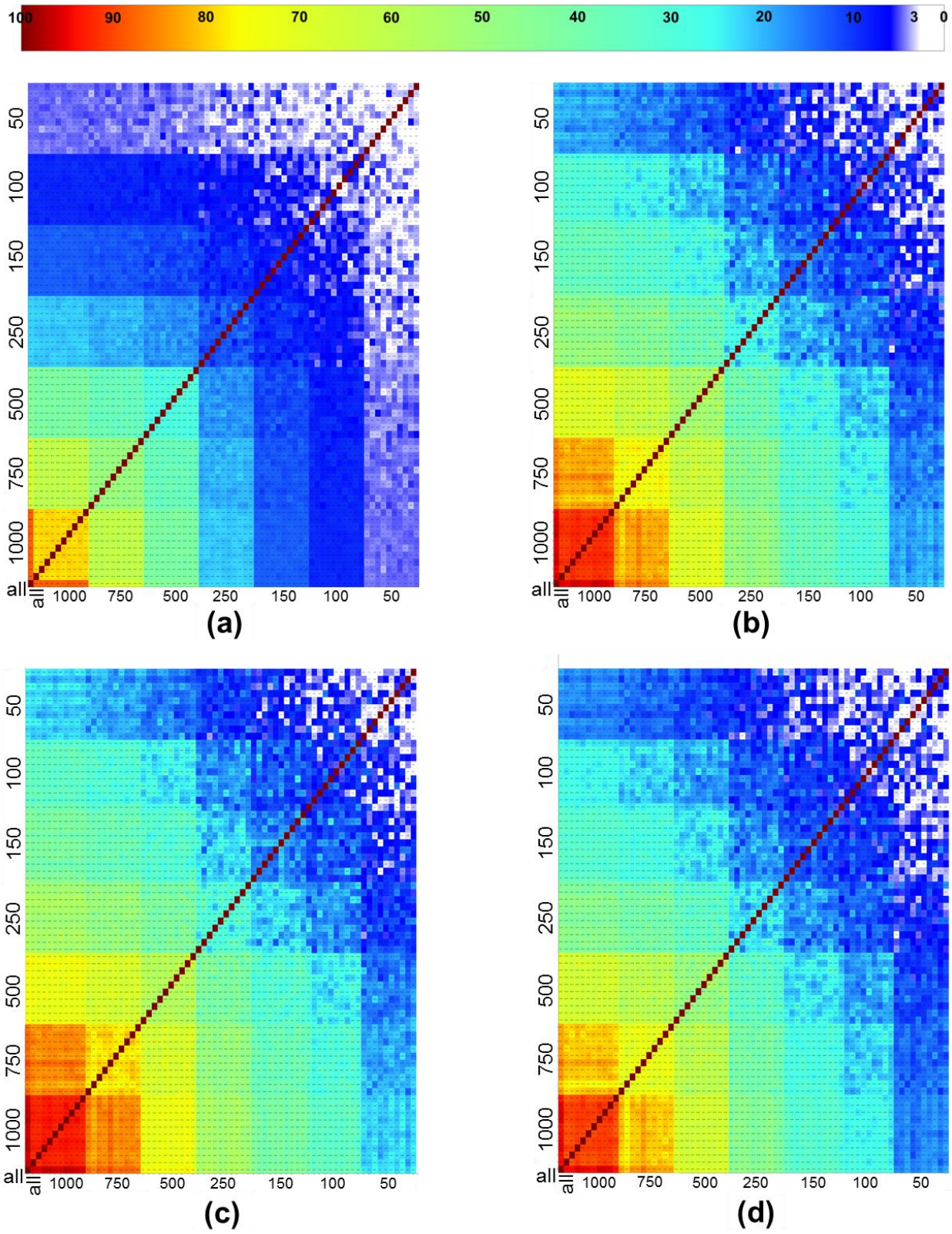


Abbildung 26.3. Ergebnisse des „Selbstähnlichkeitstestes“ für den COX2-Datensatz: (a) *ohne*, (b) bis (d) *mit* Berücksichtigung von Substruktur-Beziehungen. Einfluss der Gewichtung: (a) und (b) *gewichteter* inSARa-TSim, (c) *ungewichteter* inSARa-TSim, (d) *quadrierter Tc*.

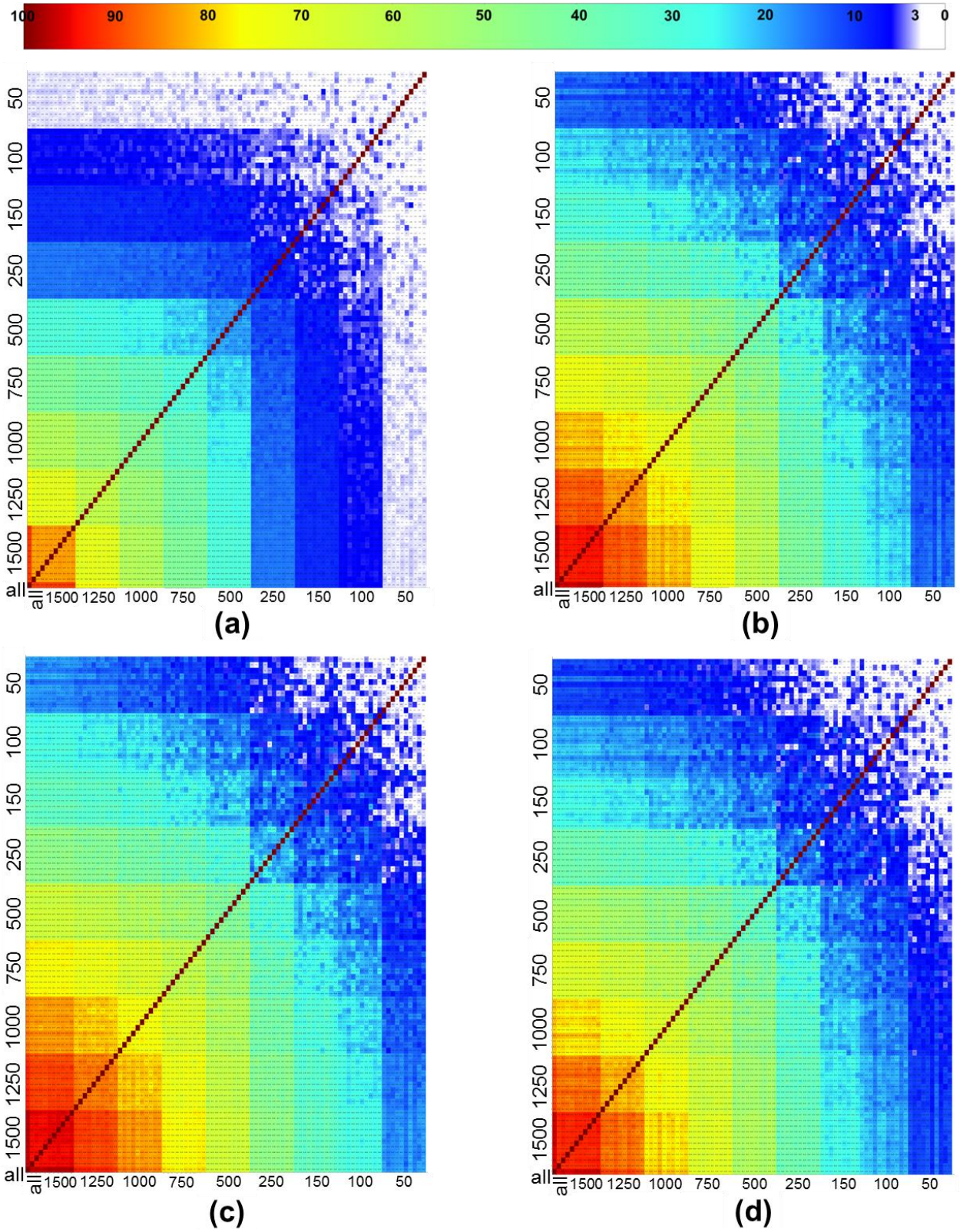


Abbildung 26.4. Ergebnisse des „Selbstähnlichkeitstestes“ für den P38-Datensatz: (a) *ohne*, (b) bis (d) *mit* Berücksichtigung von Substruktur-Beziehungen. Einfluss der Gewichtung: (a) und (b) *gewichteter* inSARa-TSim, (c) *ungewichteter* inSARa-TSim, (d) *quadrierter Tc*.

26.7. Detaillierte Ähnlichkeitskarte des ligandbasierten Target-Vergleichs

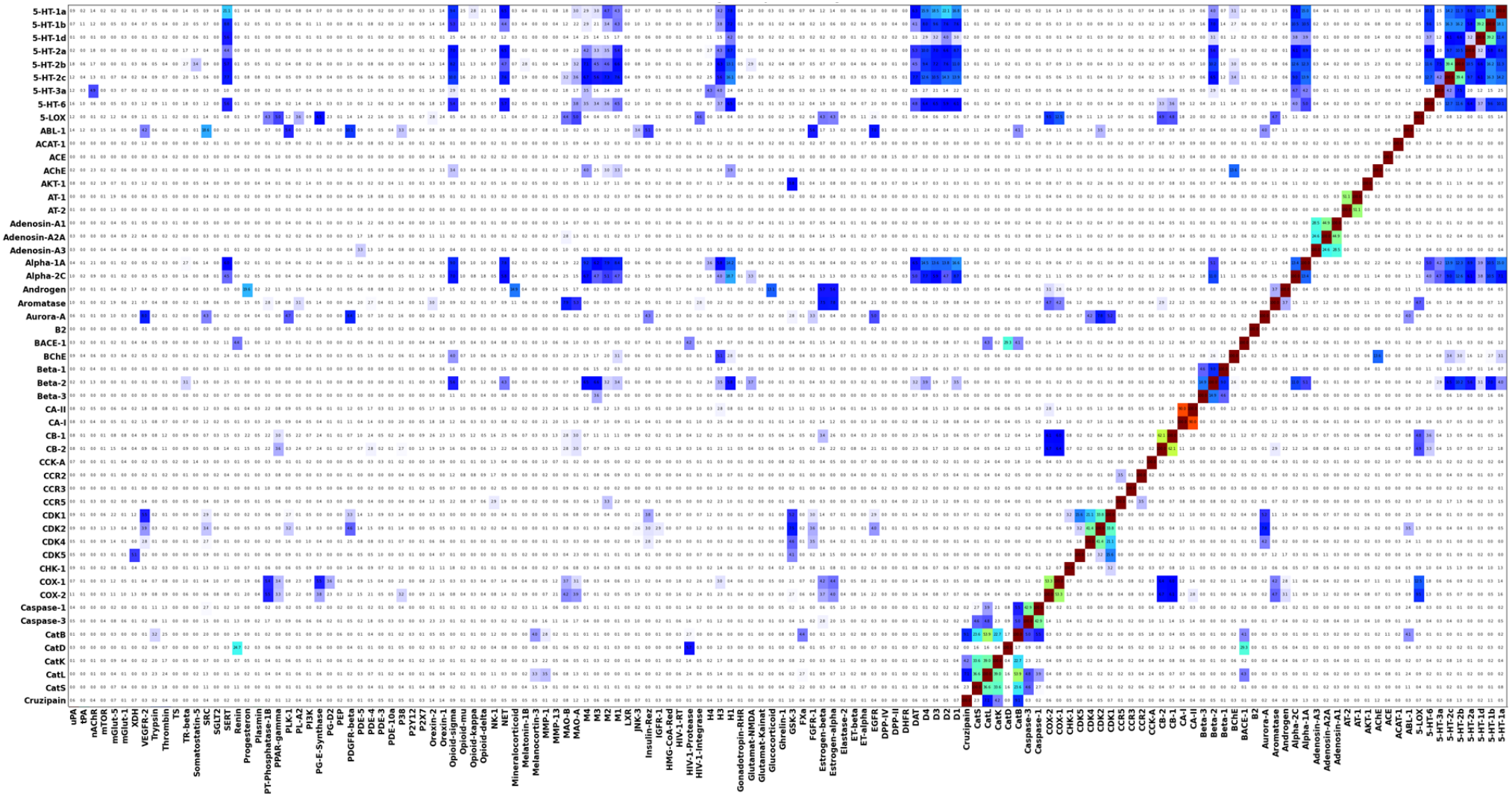


Abbildung 26.5. Teil 1 der vergrößerten Darstellung der inSARa-Netzwerk-Ähnlichkeitskarte aus Abbildung 23.5 (gewichteter inSARa-TSim).

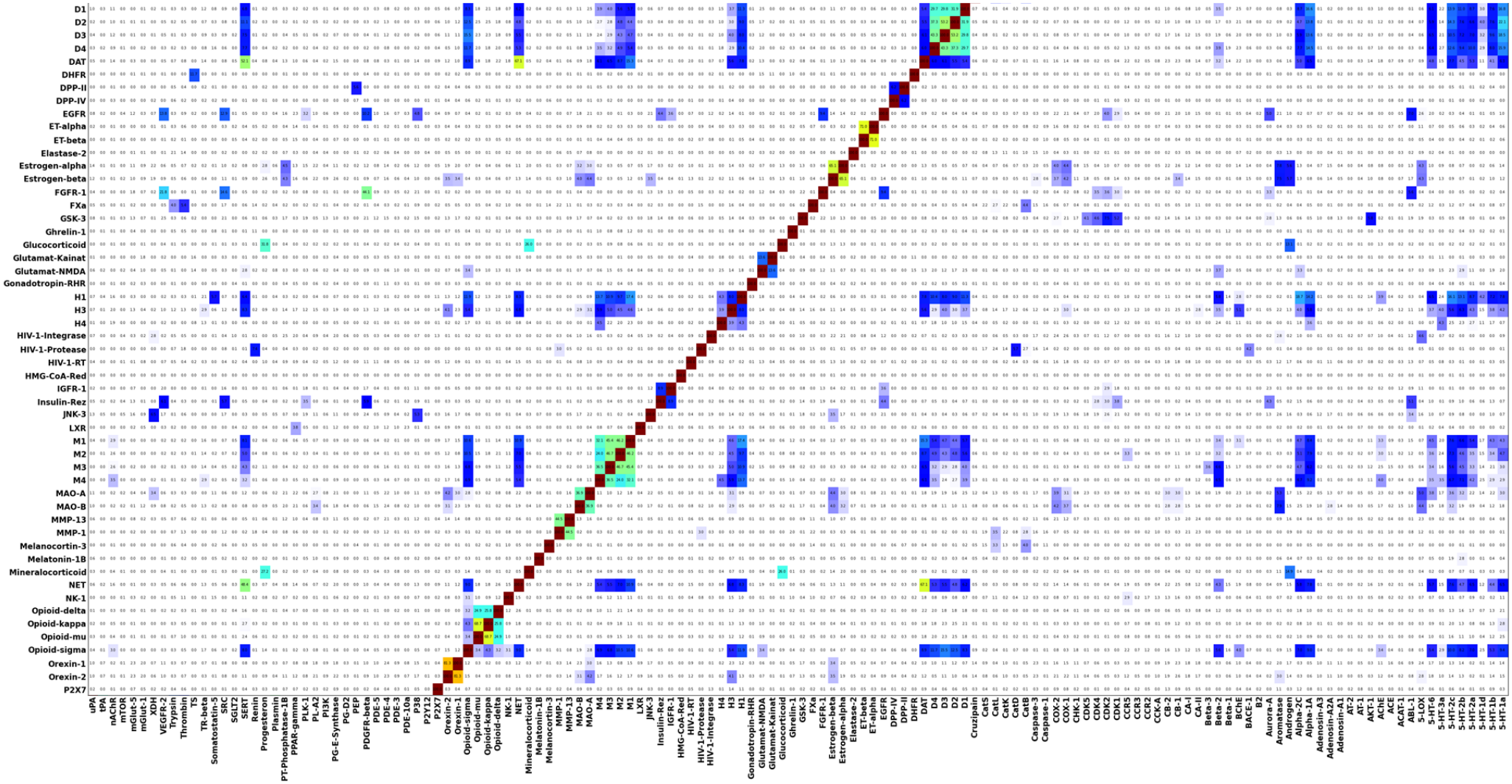


Abbildung 26.6. Teil 2 der vergrößerten Darstellung der inSARa-Netzwerk-Ähnlichkeitskarte aus Abbildung 23.5 (gewichteter inSARa-TSim).

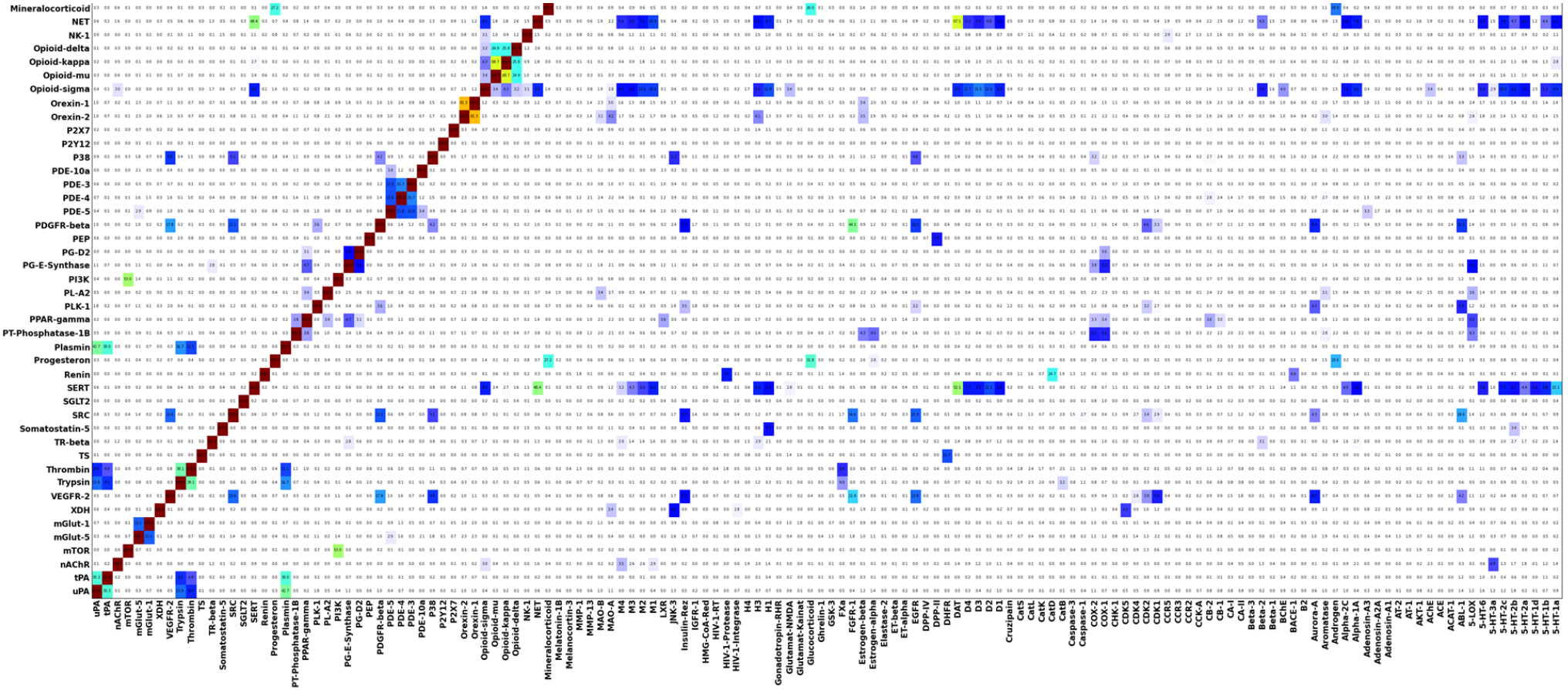


Abbildung 26.7. Teil 3 der vergrößerten Darstellung der inSARa-Netzwerk-Ähnlichkeitskarte aus Abbildung 23.5 (gewichteter inSARa-TSim).

27. Quellcode

27.1. RG-Umwandlung

RG_generation_inSARa.py

```
#!/usr/bin/python
#-*- coding: utf-8 -*-

"""
(c) Sabrina Wollenhaupt - RG-Generation for inSARa
Version 0.01 - 22-09-2013 23-44
Changes history:

"""

"""
Modulimport:
"""
import pybel #darf nicht gleichzeitig mit scipy importiert werden, sonst
Speicherzugriffsfehler!!!
from openeye.oechem import *
import cPickle as pickle
import sys, os
import time
import standard_tasks as st #eigenes Modul (Standardaufgaben)

print "\n", "Module erfolgreich importiert", "\n"



def copy_sdf_to_directory(sdf_name, start_dir, current_dir):
    """
    sdf-datei in neues Verzeichnis kopieren
    """
    print "infile: "+sdf_name+".sdf, will be copied from "+start_dir+" to new directory...\n"
    os.system("cp "+start_dir+"/"+sdf_name+".sdf "+current_dir)
    print "your file "+sdf_name+" was sucessfully copied to "+current_dir+"...\n"


def create_new_directory_for_generated_data(inputfile, date):
    """
    neues Verzeichnis erstellen
    """
    #aktuelles verzeichnis
    start_dir = os.getcwd()

    #neues verzeichnis erstellen
    new_dir = start_dir+"/"+"reduced_graphs_4_"+inputfile+"_"+date
    os.mkdir(new_dir)
    #wechseln in das neue verzeichnis
    os.chdir(new_dir)

    print "\nnew directory successfully created...\n"

    return start_dir, new_dir


def prep_mol_for_rg(mol):
    """
    Vorbereitung des Mols fuer RG-generation
    """
    OESuppressHydrogens(mol, False, False, False)
    OEFindRingAtomsAndBonds(mol)
    OEAssignAromaticFlags(mol)

    return mol


def check_for_macrocyclic_ringsystems(mol):
    """
```

```
    pruefen, dass kein Ring > 7 Atome, sonst nicht rg-umwandlung
    """
    check = 1
    for atom in mol.GetAtoms():
       if atom.IsInRing():
           if OEAtomIsInRingSize(atom, 3) == True or OEAtomIsInRingSize(atom, 4) == True or
OEAtomIsInRingSize(atom,   5)   ==   True   or   OEAtomIsInRingSize(atom,   6)   ==   True   or
OEAtomIsInRingSize(atom, 7) == True:
               pass
           else:
               check = 0
    return check


def ring_detection(mol, mollist_pybel, mol_counter):
    """
    finde  ringe  (sssr)  und  unterscheidung  aromatisch  vs  aliphatisch  und  erkennen  von
Brueckenkopfatomen
    """
    all_ring_atoms_list_oe, sssr_ring_atoms_list_oe = check_ring(mollist_pybel, mol_counter)
    aromatic_ring_atoms_list = aromatic_ring_atoms_detection(mol)
    aliphatic_ring_atoms_list     =     aliphatic_ring_atoms_detection(all_ring_atoms_list_oe,
aromatic_ring_atoms_list)
    fused_ring_atoms_list         =         fused_ring_atoms_detection(all_ring_atoms_list_oe,
sssr_ring_atoms_list_oe)
    get_atoms_coords(mol)


    ring_list = []
    ring_list.append(all_ring_atoms_list_oe)        #->[0]
    ring_list.append(sssr_ring_atoms_list_oe)       #->[1]
    ring_list.append(aromatic_ring_atoms_list)      #->[2]
    ring_list.append(aliphatic_ring_atoms_list)     #->[3]
    ring_list.append(fused_ring_atoms_list)  #->[4]

    return ring_list


def get_atoms_coords(mol):
    for atom in mol.GetAtoms():
       coords = mol.GetCoords(atom)
       print atom.GetIdx(), OEGetAtomicSymbol(atom.GetAtomicNum()), coords


def check_ring(pybel_mol_list, mol_counter):
    """
    finde mittels Pybels SSSR (Figueras, 1996) die einzelnen Ringe -> uebersetzung in OE
    """
    mol = pybel_mol_list[mol_counter]
    ring_atoms_oe_list = []
    ring_pybel_list = []
    sssr_ring_atoms_pybel_to_oe = []

    sssr = mol.sssr
    for ring in sssr:
       ring_pybel_list.append(ring)
       for ring_atom_pybel in ring._path:
           if (ring_atom_pybel-1) not in ring_atoms_oe_list:
               ring_atoms_oe_list.append(ring_atom_pybel-1)  #Pybel  startet  Mol-Idx  bei  1,  OE
bei 0
    #print ring_pybel_list
    print ring_atoms_oe_list, len(ring_atoms_oe_list)

    for ring in ring_pybel_list:
       ring_atoms_list = []
       for ring_atom_pybel in ring._path:
           ring_atoms_list.append(ring_atom_pybel-1) #Pybel startet Mol-Idx bei 1, OE bei 0
       sssr_ring_atoms_pybel_to_oe.append(ring_atoms_list)              #sssr              in
sssr_ring_atoms_pybel_to_oe: Unterliste des sssr

    print "\n", sssr_ring_atoms_pybel_to_oe, "\n"

    print "SSSR"
    counter = 1
    for sssr in sssr_ring_atoms_pybel_to_oe:
       print counter, sssr, len(sssr)
       counter += 1
```

```
    return ring_atoms_oe_list, sssr_ring_atoms_pybel_to_oe


def aromatic_ring_atoms_detection(mol):
    """
    finde alle aromatischen Ringatome
    """
    aromatic_ring_atoms_list = []
    nraromsystems, parts = OEDetermineAromaticRingSystems(mol)
    for ringidx in xrange(1, nraromsystems+1): #parts = Liste, Listenpos entspricht Atomidx,
parts_entry == 0 -> acycl., parts_entry = (1 (fuer Ringsystem 1), 2 (fuer Ringsystem 2), ...)
       for atom in mol.GetAtoms():
           if parts[atom.GetIdx()] == ringidx:
               aromatic_ring_atoms_list.append(atom.GetIdx())
    print "\naromatic ring-atoms"
    print aromatic_ring_atoms_list, len(aromatic_ring_atoms_list)

    return aromatic_ring_atoms_list


def aliphatic_ring_atoms_detection(all_ring_atoms_list, aromatic_ring_atoms_list):
    """
    finde alle aliphatischen Ringatome
    """
    aliphatic_ring_atoms_list = all_ring_atoms_list[:]

    for arom_atom in aromatic_ring_atoms_list:
       if arom_atom in aliphatic_ring_atoms_list:
           aliphatic_ring_atoms_list.remove(arom_atom)
    print "\naliphatic ring-atoms"
    print aliphatic_ring_atoms_list, len(aliphatic_ring_atoms_list)

    return aliphatic_ring_atoms_list


def fused_ring_atoms_detection(all_ring_atoms_list, sssr_ring_atoms_list):
    """
    finde die Annelierungsstellen der annelierten Ringsysteme
    """
    fused_ring_atoms_list = []

    all_sssr_atoms_list = []

    for sssr in sssr_ring_atoms_list:
       for sssr_atom in sssr:
           all_sssr_atoms_list.append(sssr_atom)

    for ring_atom in all_ring_atoms_list:
       count = all_sssr_atoms_list.count(ring_atom)
       if count > 1:
           fused_ring_atoms_list.append(ring_atom)
    print "\nfused ring-atoms"
    print fused_ring_atoms_list, len(fused_ring_atoms_list)

    return fused_ring_atoms_list


def check_feature(mol, smarts_list, feature_name, feature_atom_list):
    """
    erkenne ph4-features mittels smart-matching
    """
    general_list = []
    for smart in smarts_list:
       ss = OESubSearch(smart)
       for match in ss.Match(mol):
           for ma in match.GetAtoms():
               if ma.target.GetIdx() not in general_list: #hier im Vgl zu altem korrigiert
(statt ma not in ...)
                   general_list.append(ma.target.GetIdx())
    print feature_name, general_list

    feature_atom_list.append(general_list)

    ##zur Info
    ##hba_atom_list = feature_atom_list[0]
    ##hbd_atom_list = feature_atom_list[1]
    ##hbad_atom_list = feature_atom_list[2]
    ##pi_atom_list = feature_atom_list[3]
```

```
    ##ni_atom_list = feature_atom_list[4]

    return feature_atom_list


def check_feature_hbad(feature_atom_list, feature_name):
    """
    kombinierte hba/hbd-Eigenschaft bei einem Atom -> sinnvoll??? (wie bei Barker etc.)
    """
    hba_atom_list = feature_atom_list[0]
    hbd_atom_list = feature_atom_list[1]

    hbad_atom_list = []

    for atom in hba_atom_list:
       if atom in hbd_atom_list:
           hbad_atom_list.append(atom)
    print feature_name, hbad_atom_list

    feature_atom_list.append(hbad_atom_list)

    return feature_atom_list


def start_rg_generation(mol, ring_list, feature_atom_list):
    """
    beginn der rg-generation
    """
    all_ring_atoms_list = ring_list[0]
    print all_ring_atoms_list
    sssr_ring_atoms_list_oe = ring_list[1]
    aromatic_ring_atoms_list = ring_list[2]
    aliphatic_ring_atoms_list = ring_list[3]
    fused_ring_atoms_list = ring_list[4]

    hba_atom_list = feature_atom_list[0]
    hbd_atom_list = feature_atom_list[1]
    hbad_atom_list = feature_atom_list[2]
    pi_atom_list = feature_atom_list[3]
    ni_atom_list = feature_atom_list[4]
    black_atom_list = feature_atom_list[5]

    ##beginn mit ni-feature
    mol, ring_list, ni_pi_check = create_ph4_ni_pi_rg_nodes(mol, ni_atom_list, ring_list, 42)
    if ni_pi_check == 1:
        ##dann pi-feature
        mol, ring_list, ni_pi_check = create_ph4_ni_pi_rg_nodes(mol, pi_atom_list, ring_list,
41)

    if ni_pi_check == 1:
        ###dann ringe
        mol = create_ph4_ring_rg_nodes(mol, hba_atom_list, hbd_atom_list, hbad_atom_list,
ring_list)
        ###hba/hbd/combi-features
        mol = create_ph4_hba_hbd_rg_nodes(mol, hbad_atom_list, 29)
        mol = create_ph4_hba_hbd_rg_nodes(mol, hba_atom_list, 28)
        mol = create_ph4_hba_hbd_rg_nodes(mol, hbd_atom_list, 27)
        mol = find_linker_hydrophobic_rg_nodes(mol, black_atom_list, 30)

    return mol, ni_pi_check


def create_ph4_ni_pi_rg_nodes(mol, ph4_atom_list, ring_list, element):
    """
    umwandlung der NI bzw. PI-Eigenschaften -> berueckksichtigung von ringeinbindung
    """
    dict_group_2_atoms, dict_group_2_nbor = find_groups_and_nbors(mol, ph4_atom_list)
    mol, sssr_ring_atoms_list, ni_pi_check = generate_rg_nodes(mol, dict_group_2_atoms,
dict_group_2_nbor, ring_list, element)
    ring_list[1] = sssr_ring_atoms_list
    return mol, ring_list, ni_pi_check


def matches_out(list1, list2):
    set1 = set(list1)
    set2 = set(list2)
    set3 = set1.intersection(set2)
    list3 = list(set3)
```

```
    return list3


def generate_rg_nodes(mol, dict_group_2_atoms, dict_group_2_nbor, ring_list, element):
    """
    rg-nodes erstellen fuer NI & PI -> Sonderfall: NI/PI in Ring beruecksichtigt wie bei
Barker und Gardiner
    """
    all_ring_atoms_list = ring_list[0]
    sssr_ring_atoms_list = ring_list[1]
    fused_ring_atoms_list = ring_list[4]

    ni_pi_check = 1

    print all_ring_atoms_list

    for group_nr in dict_group_2_atoms.keys():
            group_list = dict_group_2_atoms[group_nr]
            nbor_list = dict_group_2_nbor[group_nr]

            print "group_list", group_nr, group_list
            print "nbor_list", group_nr, nbor_list
            del_list = []
            sssr_to_del_list = []
        #for atom_idx in group_list:
            #print atom_idx
            common_list = matches_out(group_list, all_ring_atoms_list)

            ##Sonderfall NI/PI-feature-atom ist Teil eines Ringes -> ganzer Ring wird als
feature umgewandelt! s.o.
            if len(common_list) >0:
                atom_idx = common_list[0]
                if atom_idx not in del_list:
                    if atom_idx not in fused_ring_atoms_list:
                        sssr_counter = 0
                        for sssr in sssr_ring_atoms_list:
                            if atom_idx in sssr:
                                for sssr_atom in sssr:
                                    if sssr_atom not in group_list and sssr_atom not in
fused_ring_atoms_list:
                                        group_list.append(sssr_atom)
                                        for atom in mol.GetAtoms():
                                            if atom.GetIdx() == sssr_atom:
                                                for nbor in atom.GetAtoms():
                                                    if nbor.GetIdx() not in nbor_list:
                                                        nbor_list.append(nbor.GetIdx())
                                                        if sssr_counter not in sssr_to_del_list:

sssr_to_del_list.append(sssr_counter)
                                                                print "\n\n"
                                                                print "SSSR", sssr_ring_atoms_list
                                                                print "sssr del", sssr_to_del_list
                                                                sssr_to_del_list.sort()
                                                                counter = 0
                                                                for sssr_nr in sssr_to_del_list:
                                                                    print "to del", sssr_nr
                                                                    ring_atoms                      =
sssr_ring_atoms_list.pop(sssr_nr-counter)
                                                                    counter += 1
                                                                print ring_atoms, "ringatoms"
                                                                for ring_atom in ring_atoms:
                                                                    if ring_atom not in group_list:
                                                                        group_list.append(ring_atom)
                                                                ##wg. Sonderfall: Ring s.o.
                                                                new_nbor_list = []
                                                                for group_atom_idx in group_list:
                                                                    print "atom", group_atom_idx,
                                                                    for mol_atom in mol.GetAtoms():
                                                                        if    mol_atom.GetIdx()    ==
group_atom_idx:
                                                                            for     group_nbor     in
mol_atom.GetAtoms():
                                                                                print          "nbor",
group_nbor.GetIdx(),
                                                                                if group_nbor.GetIdx()
not in group_list and group_nbor.GetIdx() not in new_nbor_list:

new_nbor_list.append(group_nbor.GetIdx())
```

```
                                                    #if atom in nbor_list:
                                                        #nbor_list.remove(atom)
                                                print
                                                print       "list",       group_list,
new_nbor_list

                                                ##umwandlung
                                                #loesche alle atome aus gruppe
                                                print "del",
                                                for atom in mol.GetAtoms():
                                                    if atom.GetIdx() in group_list:

del_list.append(atom.GetIdx())
                                                        print atom.GetIdx(),
                                                        mol.DeleteAtom(atom)
                                                print "\n"
                                                #erstelle neuen rg-node
                                                node = mol.NewAtom(element)
                                                #verbinde   mit   nachbar-atomen   aus
nbor_list
                                                print "nbor"
                                                for atom in mol.GetAtoms():
                                                    if        atom.GetIdx()        in
new_nbor_list:
                                                        print atom.GetIdx(),
                                                        edge    =    mol.NewBond(node,
atom, 1)
                                                print "\n"
                            sssr_counter += 1
                    #else:
                        #print "feature-atom (NI/PI) gehoert zu 2 Ringen (fused_ring_atom)!!! ->
Abbruch, da kein Vorgehen implementiert ist..."
                        ##sys.exit(3)
                        #ni_pi_check = 0
                        #return mol, sssr_ring_atoms_list, ni_pi_check
                else:
                    atom_idx = group_list[0]
                    if atom_idx not in del_list:
                        new_nbor_list = []
                        for group_atom_idx in group_list:
                            print "atom", group_atom_idx,
                            for mol_atom in mol.GetAtoms():
                                if mol_atom.GetIdx() == group_atom_idx:
                                    for group_nbor in mol_atom.GetAtoms():
                                        print "nbor", group_nbor.GetIdx(),
                                        if    group_nbor.GetIdx()    not    in    group_list    and
group_nbor.GetIdx() not in new_nbor_list:
                                            new_nbor_list.append(group_nbor.GetIdx())
                        print
                        print "list", group_list, new_nbor_list

                        ##umwandlung
                        #loesche alle atome aus gruppe
                        print "del",
                        for atom in mol.GetAtoms():
                            if atom.GetIdx() in group_list:
                                del_list.append(atom.GetIdx())
                                print atom.GetIdx(),
                                mol.DeleteAtom(atom)
                        print "\n"
                        #erstelle neuen rg-node
                        node = mol.NewAtom(element)
                        #verbinde mit nachbar-atomen aus nbor_list
                        print "nbor"
                        for atom in mol.GetAtoms():
                            if atom.GetIdx() in new_nbor_list:
                                print atom.GetIdx(),
                                edge = mol.NewBond(node, atom, 1)
                        print "\n"

    return mol, sssr_ring_atoms_list, ni_pi_check


def find_groups_and_nbors(mol, ph4_atom_list):
    """
    gruppiere benachbarte feature-atome und finde nachbar-atome
    """
```

```
dict_group_2_nbor = {}
dict_group_2_atoms = {}
dict_atom_2_group_nr = {}
used_list = []
group_nr = -1
for atom in mol.GetAtoms():
    if atom.GetIdx() in ph4_atom_list:
        #print
        #print atom.GetIdx()
        #print used_list
        if atom.GetIdx() not in used_list:
            nbor_list = []
            group_list = []
            group_list.append(atom.GetIdx())
            group_nr += 1
            used_list.append(atom.GetIdx())
            dict_atom_2_group_nr[atom.GetIdx()] = group_nr
            #print "group_nr", group_nr

            for nbor in atom.GetAtoms():
                #print nbor.GetIdx(),
                if nbor.GetIdx() not in ph4_atom_list:
                    #print "anderer nachbar",
                    if nbor.GetIdx() not in nbor_list:
                          nbor_list.append(nbor.GetIdx())
                else:
                    #pruefen ob zur gleichen gruppe = gleicher typ gehoert
                    if nbor.GetIdx() not in group_list:
                        group_list.append(nbor.GetIdx())
                        used_list.append(nbor.GetIdx())
                        dict_atom_2_group_nr[nbor.GetIdx()] = group_nr
            #print used_list
            dict_group_2_atoms[group_nr] = group_list
            dict_group_2_nbor[group_nr] = nbor_list

        else:
            group = dict_atom_2_group_nr[atom.GetIdx()]
            #print group
            group_list = dict_group_2_atoms[group]
            nbor_list = dict_group_2_nbor[group]

            for nbor in atom.GetAtoms():
                #print nbor.GetIdx(),
                if nbor.GetIdx() not in ph4_atom_list:
                    #print "anderer nachbar",
                    if nbor.GetIdx() not in nbor_list:
                          nbor_list.append(nbor.GetIdx())
                else:
                    if nbor.GetIdx() not in group_list:
                        group_list.append(nbor.GetIdx())
                        used_list.append(nbor.GetIdx())
                        dict_atom_2_group_nr[nbor.GetIdx()] = group
            #print used_list
            dict_group_2_atoms[group] = group_list
            dict_group_2_nbor[group] = nbor_list
#print
#for k,v in dict_group_2_atoms.iteritems():
    #print k,v
#for k,v in dict_group_2_nbor.iteritems():
    #print k,v
#for k,v in dict_atom_2_group_nr.iteritems():
    #print k,v

dict_same_group = {}
for k,v in dict_group_2_atoms.iteritems():
    for v1 in v:
        for p,m in dict_group_2_atoms.iteritems():
            if p!=k:
                if v1 in m:
                    if k>p:
                        if k not in dict_same_group:
                            dict_same_group[k] = [p]
                        else:
                            if p not in dict_same_group[k]:
                                dict_same_group[k].append(p)
                    else:
                          if p not in dict_same_group:
                              dict_same_group[p] = [k]
```

```
                            else:
                                if k not in dict_same_group[p]:
                                    dict_same_group[p].append(k)
    print "\n"
    for k,v in dict_same_group.iteritems():
        #print k,v
        for v1 in v:
            #print v1
            if dict_group_2_atoms[k] != []:
                #print dict_group_2_atoms[v1]
                for value in dict_group_2_atoms[v1]:
                    #print "value", value
                    if value not in dict_group_2_atoms[k]:
                        #print dict_group_2_atoms[k]
                        dict_group_2_atoms[k].append(value)
                        #print dict_group_2_atoms[k]
                    #dict_group_2_atoms[v1].remove(value)
    for k,v in dict_same_group.iteritems():
        for v1 in v:
            dict_group_2_atoms[v1] = []


    new_dict_group_2_atoms = {}
    counter = 0
    for k,v in dict_group_2_atoms.iteritems():
        if v != []:
            new_dict_group_2_atoms[counter] = v
            counter += 1

    new_dict_group_2_nbor = {}
    for k,v in new_dict_group_2_atoms.iteritems():
        print "groups", k,v
        for v1 in v:
            if k not in new_dict_group_2_nbor:
                new_dict_group_2_nbor[k] = []
            for atom in mol.GetAtoms():
                if atom.GetIdx() == v1:
                    for nbor in atom.GetAtoms():
                        if nbor.GetIdx() not in v and nbor.GetIdx() not in
new_dict_group_2_nbor[k]:
                            new_dict_group_2_nbor[k].append(nbor.GetIdx())

    for k,v in new_dict_group_2_nbor.iteritems():
        print "nbors", k,v


    return new_dict_group_2_atoms, new_dict_group_2_nbor

def find_linker_hydrophobic_rg_nodes(mol, allowed_atom_list, element):
    linker_list = []
    atoms_2_del_list = []
    for atom in mol.GetAtoms():
        if atom.GetAtomicNum() not in superatom_atomicnum_harper_list:
            if atom.GetIdx() in allowed_atom_list:
                linker_list.append(atom.GetIdx())
            else:
                atoms_2_del_list.append(atom.GetIdx())

    print "linker", linker_list
    print "to del", atoms_2_del_list
    mol = create_ph4_hba_hbd_rg_nodes(mol, linker_list, element)
    #mol = entferne_nicht_umgewandelte_atome(mol, atoms_2_del_list)
    return mol


def entferne_nicht_umgewandelte_atome(mol, atoms_2_del_list):
    dict_group_2_atoms, dict_group_2_nbor = find_groups_and_nbors(mol, atoms_2_del_list)
    mol = entferne_ueberfluessige_atome(mol, dict_group_2_atoms, dict_group_2_nbor)


def entferne_ueberfluessige_atome(mol, dict_group_2_atoms, dict_group_2_nbor):
    pass


def create_ph4_hba_hbd_rg_nodes(mol, ph4_atom_list, element):
    dict_group_2_atoms, dict_group_2_nbor = find_groups_and_nbors(mol, ph4_atom_list)
    mol = generate_rg_nodes_hb(mol, dict_group_2_atoms, dict_group_2_nbor, element)
```

```
    return mol


def generate_rg_nodes_hb(mol, dict_group_2_atoms, dict_group_2_nbor, element):
    """
    rg-nodes erstellen fuer HBAD/HBA/HBD
    """

    for group_nr in dict_group_2_atoms.keys():
        group_list = dict_group_2_atoms[group_nr]
        nbor_list = dict_group_2_nbor[group_nr]

        ##umwandlung
        #loesche alle atome aus gruppe
        for atom in mol.GetAtoms():
            if atom.GetIdx() in group_list:
                mol.DeleteAtom(atom)
        #erstelle neuen rg-node
        node = mol.NewAtom(element)
        #verbinde mit nachbar-atomen aus nbor_list
        for atom in mol.GetAtoms():
            if atom.GetIdx() in nbor_list:
                edge = mol.NewBond(node, atom, 1)
    return mol



def check_hb_features_in_ring(sssr, hba_atom_list, hbd_atom_list):
    hba_marker = 0
    hbd_marker = 0

    for sssr_atom in sssr:
        if hba_marker == 0 or hbd_marker == 0:
            if hba_marker == 0:
                if sssr_atom in hba_atom_list:
                    hba_marker = 1
            if hbd_marker == 0:
                if sssr_atom in hbd_atom_list:
                    hbd_marker = 1
        else:
            break

    return hba_marker, hbd_marker


def create_ph4_ring_rg_nodes(mol, hba_atom_list, hbd_atom_list, hbad_atom_list, ring_list):
    """
    umwandlung der ringe mit beruecksichtigung der hba/hbd-eigenschaften
    """
    all_ring_atoms_list = ring_list[0]
    sssr_ring_atoms_list = ring_list[1]
    aromatic_ring_atoms_list = ring_list[2]
    aliphatic_ring_atoms_list = ring_list[3]
    fused_ring_atoms_list = ring_list[4]

    aromatic_elements_list = [21,22,23,24]
    aliphatic_elements_list = [72,73,74,75]

    sssr_counter = 0
    dict_ring_nr_2_element = {}
    dict_ring1_2_ring2 = {}

    for sssr in sssr_ring_atoms_list:

        #bestimmen ob hba/d-eigenschaften zusaetzlich im ring
        hba_marker, hbd_marker = check_hb_features_in_ring(sssr, hba_atom_list, hbd_atom_list)

        print "marker", sssr, hba_marker, hbd_marker

        #aromatischer ring
        if sssr[0] in aromatic_ring_atoms_list:
            if hba_marker == 1 or hbd_marker == 1:
                if hba_marker == 1 and hbd_marker == 1:
                    element = aromatic_elements_list[3]
                elif hba_marker == 1 and hbd_marker == 0:
                    element = aromatic_elements_list[2]
                elif hba_marker == 0 and hbd_marker == 1:
                    element = aromatic_elements_list[1]
```

```
            else:
                element = aromatic_elements_list[0]

        #aliphatischer ring
        elif sssr[0] in aliphatic_ring_atoms_list:
            if hba_marker == 1 or hbd_marker == 1:
                if hba_marker == 1 and hbd_marker == 1:
                    element = aliphatic_elements_list[3]
                elif hba_marker == 1 and hbd_marker == 0:
                    element = aliphatic_elements_list[2]
                elif hba_marker == 0 and hbd_marker == 1:
                    element = aliphatic_elements_list[1]
            else:
                element = aliphatic_elements_list[0]

        else:
            print "Fehler bei Ringtyp-Erkennung -> bitte ueberpruefen!"
            sys.exit(4)

        dict_ring_nr_2_element[sssr_counter] = element
        sssr_counter += 1

    for fused_ring_atom in fused_ring_atoms_list:
        kv_list = []
        for x in range(0, len(sssr_ring_atoms_list)):
            if fused_ring_atom in sssr_ring_atoms_list[x]:
                kv_list.append(x)
                if len(kv_list) == 2:
                    key = min(kv_list)
                    value = max(kv_list)
                    if key not in dict_ring1_2_ring2:
                        dict_ring1_2_ring2[key] = [value]
                    else:
                        if value not in dict_ring1_2_ring2[key]:
                            dict_ring1_2_ring2[key].append(value)

    sssr_counter = 0
    dict_ring_nr_2_idx = {}

    edge_list = []
    for sssr in sssr_ring_atoms_list:
        nbor_list = []
        #anknuepfungspunkte suchen / Ausnahme Brueckenkopfatome (-> bleiben erstmal erhalten)
        for atom in mol.GetAtoms():
            if atom.GetIdx() in sssr:
                for nbor in atom.GetAtoms():
                    if  nbor.GetIdx()  not  in  sssr  and  nbor.GetIdx()  not  in  nbor_list  and
nbor.GetIdx() not in fused_ring_atoms_list:
                        nbor_list.append(nbor.GetIdx())

        #loeschen aller ringatome
        for atom in mol.GetAtoms():
            if atom.GetIdx() in sssr:
                mol.DeleteAtom(atom)

        #erstelle neuen ring-node
        element = dict_ring_nr_2_element[sssr_counter]
        node = mol.NewAtom(element)
        new_idx = node.GetIdx()
        dict_ring_nr_2_idx[sssr_counter] = new_idx
        sssr_counter += 1

        #verbinde mit nachbar-atomen aus nbor_list
        for atom in mol.GetAtoms():
            if atom.GetIdx() in nbor_list:
                edge = mol.NewBond(node, atom, 1)
                edge_list.append(edge)

    #ring-annelierung -> Doppelbdg
    if fused_ring_atoms_list != []:
        for k,v in dict_ring1_2_ring2.iteritems():
            atom_idx_1 = dict_ring_nr_2_idx[k]
            for bond in edge_list:
                if bond.GetBgnIdx() < bond.GetEndIdx():
                    for vx in v:
                        atom_idx_2 = dict_ring_nr_2_idx[vx]
                        if atom_idx_1 == bond.GetBgnIdx() and atom_idx_2 == bond.GetEndIdx():
                            bond.SetOrder(2)
```

```
                        print "fused_ring"
            else:
                for vx in v:
                    atom_idx_2 = dict_ring_nr_2_idx[vx]
                    if atom_idx_2 == bond.GetBgnIdx() and atom_idx_1 == bond.GetEndIdx():
                        bond.SetOrder(2)
                        print "fused_ring"

    return mol


def check_for_correct_rg_encoding(rg_mol):
    check = 1
    for atom in rg_mol.GetAtoms():
        if atom.GetAtomicNum() not in superatom_atomicnum_harper_list:
            check = 0
    return check

def check_for_molecule_splitting(mol):
    check = 1
    count, parts = OEDetermineComponents(mol)
    if count > 1:
        check = 0
    return check

##=========================SMARTS-
pattern============================================================================##

#negativ ionisierbar: Saueren(C,S,P), Tetrazol, Imide, negative Ladung nicht benachbart zu
positiver Ladung...
ni_list  =  ["[C,S,P](=[O,S])(-[O;H1,-1])",  "[S](=[O,S])(=[O,S])(-[O;H1,-1])",  "[P](=[O,S])(-
[O;H1,-1])(-[O;H1,-1])", "[n1nnnc1]", "[n1nncn1]",
           "[*-;!$([*-]([*+]))]",   "[N;H1,-1](S(=O)(=O))(C(=O))",   "[N;H1,-1](C(=O))(C(=O))",
"[N;H1,-1](S(=O)(=O))(S(=O)(=O))", "[N;H1,-1](C(F)(F)(F))(S(=O)(=O))"]

#positiv ionisierbar: aliphatische (prim, sek, tert) Amine, Guanidin, Amidin, positive Ladung
nicht adjazent zu neg Ladung(umfasst Nitro & quat N)...
pi_list                          =                          ["[$([N;H2&+0,H3&+1][C;!$([C](=*))])]",
"[$([N;H1&+0,H2&+1]([C;!$([C](=*))])[C;!$([C](=*))])]",
"[$([N;H1&+0,H2&+1]([C;!$([C](=*))])([C;!$([C](=*))])[C;!$([C](=*))])]",
           "[N;X2&+0,X3&+1&!H0](=C([NX3])([NX3]))",           "[N;X2&+0,X3&+1&!H0](=C([NX3]))",
"[N;v3X3,v4X4+][CX3](=[N;v3X2,v4X3+])[N;v3X3,v4X4+]", "[*+;!$([*+]([*-]))]"]


#HBA: #vgl. Gillet "similarity searching rgs" plus erweiterung um nitro
#plus (-23-09-12-) erweiterung um Ausschluss von Amid/Sulfonamid etc. N-Atomen (vgl. Pharao
u.a.) -> ;!$([N]C(=O));!$([N][S,P](=O)(=O))]
#plus (-23-09-12-) erweiterung um Ausschluss von Amidin-Guanidin etc. N-Atomen (-keine pos.
Ladung erlaubt, delokal. e-Paar) -> ;!$([N]C(=N))]
#plus (-24-09-12-) erweiterung damit Carbonyl-C und Sulfonyl-S etc mit zu Acc gezaehlt werden
-> vereinfacht umwandlung "[$([C,S,P]=[N,S,O,P])]"
hba_list                                                                                    =
["[!#6;!$([+1,+2,+3]);!$([F,Cl,Br,I]);!$([#1]~[#6])!$([o,s,nX3]);!$([Nv5,Pv5,Sv4,Sv6]);!$([N]C
(=O));!$([N][S,P](=O)(=O));!$([N]C(=N))]",                          "[NX3+1](-[OX1-1])(=[OX1+0])",
"[$([C,S,P]=[N,S,O,P])]",  "[$(C#[N])]"]  #[NX3+1](-[OX1-1])(=[OX1+0])  =  Nitro:  wg  N,  sonst
nicht umgewandelt

#HBD:
hbd_list = ["[!#6;!H0;!#1]"]


#isolating carbon:
ic_list     =     ["[$([C!R2;C!R3]);!$(C=[N,S,O,P]);!$(C#[N,S,O,P])]"]     #erfasst     auch
Ethylengruppen(!), sinnvoll(?)

#black_list => Ausschluss von Carbonyl-C oder Sulfonyl-S als linker-feature
black_list = ["[!$([C,P,S,N]=[N,S,O,P]);!$(C#[N,S,O,P])]"]


#hydrophobe Endgruppen: Halogene (keine Imidbeteiligung), non-HB-Atome, auch CHF2 oder CF3,
CFH2 etc.
hydrophobic_ends_list  =  ["[$([F,Cl,Br,I]);!$(F(C(F)(F)[N;H1,-1](S(=O)(=O))))]",  "[$([CX4](-
[F,Cl,Br,I])(-[F,Cl,Br,I]))]", "[$([CX4](-[F,Cl,Br,I]))]"]
```

```
##=================================================================================================
===============##

superatom_code_harper_list = ['Sc', 'Ti', 'V', 'Cr', 'Mn', 'Fe', 'Hf', 'Ta', 'W', 'Re', 'Y',
'Zr', 'Co', 'Ni', 'Cu', 'Nb', 'Mo', 'Zn', 'Hg']

superatom_description_harper_list = ['aromatic ring nonfeature', 'aromatic ring donor',
'aromatic ring acceptor',
'aromatic ring donor & acceptor', 'aromatic ring positively ionizable', 'aromatic ring
negatively ionizable',
'aliphatic ring nonfeature', 'aliphatic ring donor', 'aliphatic ring acceptor', 'aliphatic
ring donor & acceptor',
'aliphatic ring positively ionizable', 'aliphatic ring negatively ionizable', 'feature node
donor', 'feature node acceptor',
'feature node donor & acceptor', 'feature node positively ionizable', 'feature node negatively
ionizable', 'link node', 'hydrophobic endcaping']

superatom_atomicnum_harper_list = [21, 22, 23, 24, 25, 26, 72, 73, 74, 75, 39, 40, 27, 28, 29,
41, 42, 30, 80]

harper_atomicnum_superatom_dict          =          dict(zip(superatom_atomicnum_harper_list,
superatom_code_harper_list))



##================================Beginn                                                          des
Hauptprogrammes=====================================================##

#start-zeit
start, date = st.note_start_time()

start_dir             = os.getcwd()
all_files             = os.listdir(start_dir)

all_mol_files = [name for name in all_files if 'sdwash_sdfilter_duplfrei_fps_added_fp2.sdf'
in name]

for mol_file in all_mol_files:

    mol_counter = 0
    error_counter = 0
    rg_mollist = []

    dateiname = mol_file.split('.')[0]

    #einlesen der mols (oe)
    mollist = st.einlesen_der_mols(dateiname)

    #einlesen der mols (pybel)
    mollist_pybel = st.einlesen_der_mols_pybel(dateiname)


    #neues Verzeichnis erstellen
    start_dir, current_dir = create_new_directory_for_generated_data(dateiname, date)

    copy_sdf_to_directory(dateiname, start_dir, current_dir)


    #log_datei oeffnen
    log_datei = open('log_datei_'+dateiname+'.txt', 'w')
    result_datei = open('result_datei_'+dateiname+'.txt', 'w')

    print >> result_datei, "RG-generation started: ", start, "...\n"
    print   >>   result_datei,   "your   file   "+dateiname+"   was   sucessfully   copied   to
"+current_dir+"...\n"

    atom_check = 1
    ring_check = 1
    component_check = 1
    ni_pi_check = 1

    splitted_mols = []
    splitted_rgs = []

    #jedes mol durchgehen
    for mol in mollist:

        feature_atom_list = []
```

```
        ##nouveau: 16-6-12 => sd-tag von original-mol auslesen -> spaeter auf RG uebertragen
        ##wird nicht gebraucht, da altes mol benutzt wird -> aber unten doch!
        dict_sdtag_sdvalue = st.get_all_sdtags_from_mol(mol)

        #Vorbereitung fuer RG-generation
        mol = prep_mol_for_rg(mol)

        #Kontrollieren, ob mol prozessiert werden kann: Ringgroesse max. 7
        ring_check = check_for_macrocyclic_ringsystems(mol)

        #ausgabe
        print "\n------------------------------"
        print "molecule_nr", mol_counter+1, mol.GetTitle(), "\n"
        print "\n\n"

        #ring_check = 1 -> RG-gen. moeglich
        if ring_check == 1:

            ring_list = ring_detection(mol, mollist_pybel, mol_counter)
            feature_atom_list = check_feature(mol, hba_list, "hba", feature_atom_list)
            feature_atom_list = check_feature(mol, hbd_list, "hbd", feature_atom_list)
            feature_atom_list = check_feature_hbad(feature_atom_list, "hbad")
            feature_atom_list = check_feature(mol, pi_list, "pi", feature_atom_list)
            feature_atom_list = check_feature(mol, ni_list, "ni", feature_atom_list)
            feature_atom_list     =     check_feature(mol,     black_list,     "allowed_atoms",
feature_atom_list)

            mol, ni_pi_check = start_rg_generation(mol, ring_list, feature_atom_list)

            atom_check = check_for_correct_rg_encoding(mol)

            component_check = check_for_molecule_splitting(mol)


            ##result_datei
            print >> result_datei, "\n------------------------------"
            print >> result_datei, "molecule_nr", mol_counter+1, mol.GetTitle(), "\n"
            print >> result_datei, "\n\n"

            if atom_check == 1 and component_check == 1 and ni_pi_check == 1:
                #Ausgabe der RGs inkl. SD-tag!!!
                print "mol successfully processed..."
                rg_mollist.append(mol)
            else:
                if ni_pi_check == 0:
                    print "failure: failure while rg-generation, ni_pi_failure found...\n"
                    print >> result_datei, "failure: failure while rg-generation, ni_pi_failure
found...\n"
                    #log_datei
                    print >> log_datei, "\n------------------------------"
                    print >> log_datei, "molecule_nr", mol_counter+1, mol.GetTitle(), "\n"
                    print >> log_datei, "\n\n"
                    print >> log_datei, "failure: failure while rg-generation, ni_pi_failure
found...\n"
                elif atom_check == 0:
                    print "failure: failure while rg-generation, unknown atom-type found...\n"
                    print >> result_datei, "failure: failure while rg-generation, unknown atom-
type found...\n"
                    #log_datei
                    print >> log_datei, "\n------------------------------"
                    print >> log_datei, "molecule_nr", mol_counter+1, mol.GetTitle(), "\n"
                    print >> log_datei, "\n\n"
                    print >> log_datei, "failure: failure while rg-generation, unknown atom-
type found...\n"
                elif component_check == 0:
                    print "failure: failure while rg-generation, splitted RG found...\n"
                    print >> result_datei, "failure: failure while rg-generation, splitted RG
found...\n"
                    #log_datei
                    print >> log_datei, "\n------------------------------"
                    print >> log_datei, "molecule_nr", mol_counter+1, mol.GetTitle(), "\n"
                    print >> log_datei, "\n\n"
                    print >> log_datei, "failure: failure while rg-generation, splitted RG
found...\n"
            if component_check == 0:
                  isosmi = OEGetSDData(mol, "ISOCANSMI")
                  mol_k = OEGraphMol()
```

```
                OEParseSmiles(mol_k, isosmi)
                for tag_name, tag_content in dict_sdtag_sdvalue.iteritems():
                    st.add_sdtag_to_mol(mol_k, tag_name, tag_content)
                splitted_mols.append(mol_k)
                splitted_rgs.append(mol)
        else:
            print "\n------------------------------"
            print "molecule_nr", mol_counter+1, mol.GetTitle(), "\n"
            print "\n\n"
            print "failure: mol can not be processed due to ring-size > 7...\n"
            #result_datei
            print >> result_datei, "\n------------------------------"
            print >> result_datei, "molecule_nr", mol_counter+1, mol.GetTitle(), "\n"
            print >> result_datei, "\n\n"
            print >> result_datei, "failure: mol can not be processed due to ring-size >
7...\n"
            #log_datei
            print >> log_datei, "\n------------------------------"
            print >> log_datei, "molecule_nr", mol_counter+1, mol.GetTitle(), "\n"
            print >> log_datei, "\n\n"
            print >> log_datei, "failure: mol can not be processed due to ring-size > 7...\n"

        mol_counter += 1

        if atom_check == 0 or ring_check == 0 or ni_pi_check == 0 or component_check == 0:
            error_counter += 1

    st.write_mols_to_sdf(dateiname+"_rgs", rg_mollist)
    st.write_mols_to_sdf(dateiname+"_splitted_mols", splitted_mols)
    st.write_mols_to_sdf(dateiname+"_splitted_rgs", splitted_rgs)


    end, enddate = st.note_end_time()



    print "\n--------------------------------\n"
    print "Gesamtstatistik:\n"
    print "Fehler:", error_counter, "von", mol_counter, "Gesamt-molecules\n\n"

    ##result_datei
    print >> result_datei, "\n--------------------------------\n"
    print >> result_datei, "RG-generation ended: ", end, "...\n"
    print >> result_datei, "\n--------------------------------\n"
    print >> result_datei, "Gesamtstatistik:\n"
    print   >>   result_datei,   "Fehler:",   error_counter,   "von",   mol_counter,   "Gesamt-
molecules\n\n"
    #log_datei
    print >> log_datei, "\n--------------------------------\n"
    print >> log_datei, "RG-generation started: ", start, "...\n"
    print >> log_datei, "RG-generation ended: ", end, "...\n"
    print >> log_datei, "\n--------------------------------\n"
    print >> log_datei, "Gesamtstatistik:\n"
    print >> log_datei, "Fehler:", error_counter, "von", mol_counter, "Gesamt-molecules\n\n"

    log_datei.close()
    result_datei.close()

    os.chdir(start_dir)
```

27.2. MCS-Berechnung

inSARa_MCS_generation_mit_Kommandozeilenfunktion.py

```
#!/usr/bin/python
#-*- coding: utf-8 -*-
"""
(c) Sabrina Wollenhaupt - inSARa for SAR-Analysis (MCS-Berechnung)
MCS-Berechnung ausgelagert von Netzwerk-Erzeugung -> nur einmal durchfuehren und dann nur noch
pkl-Datei laden

Version 0.01 - 10-07-2013 21-04
Changes history:
```

```
-10-07-13 21:07: Option fuer Variation der MCS-Min-Groesse
-10-07-13 21:07: ZINC-black-list optional einschlieÃŸen
-10-07-13 21:07: Optionen fuer Kommandozeilen-Funktion (optparse)
-20-10-13: Option fuer Algorithmus-Wahl
-20-10-13: RG-Duplikate in MCS-Pool via Diagonale in MCS-Matrix
-20-10-13:  pruefen,  welche  Dateien  schon  verfuegbar  ->  fuer  Beschleunigung  weiterer
Berechnungen nutzen
-26-11-13: Bug: remove(z)
"""

"""
Modulimport
"""

from openeye.oechem import *
import cPickle as pickle
import sys, os
import networkx as nx
import standard_tasks as st #eigenes Modul (Standardaufgaben)
from classes_for_insara_3 import * #eigenes Modul fuer eigene Klassen...

from optparse import OptionParser ##Optionen fuer Kommandozeilen-Funktion

print "\n", "Module in 'inSARa MCS' erfolgreich importiert", "\n"



def check_rg_duplicate(mol_single_list):
    """
    sucht  nach  Duplikaten  bei  den  RGs  und  erstellt  eine  Liste  einzigartiger  RGs
(rg_duplfrei_liste) -> um Effizienz der MCS-Rechnungen zu steigern
    -> merken, welche RGs-Duplikate -> auf Diagonale der MCS-Matrix als MCS speichern!
    """
    check_liste_duplicate = []
    dict_rg_cansmi_titles = {}
    dict_rg_cansmi_duplfrei_title = {}
    duplfrei_name_counter = 0
    rgs_mit_duplikaten = [] #in mcs-matrix beruecksichtigen

    dict_duplfrei_title_2_mcs_pos = {}

    rg_duplfrei_liste = [] #enthaelt rgmols (ohne Duplikate)

    for ms in mol_single_list:
       if ms.rg_cansmi not in check_liste_duplicate:
           check_liste_duplicate.append(ms.rg_cansmi)
           dict_rg_cansmi_titles[ms.rg_cansmi] = [ms.title]

           duplfrei_name = "duplfrei_"+str(duplfrei_name_counter)
           dict_duplfrei_title_2_mcs_pos[duplfrei_name] = duplfrei_name_counter
           Mol_Single.set_duplfrei_name(ms, duplfrei_name)
           Mol_Single.set_duplfrei_mol_pos_nr(ms, duplfrei_name)

           rgmol = OEGraphMol()
           OEParseSmiles(rgmol, ms.rg_cansmi)
           rg_duplfrei_liste.append(rgmol)

           dict_rg_cansmi_duplfrei_title[ms.rg_cansmi] = duplfrei_name
           duplfrei_name_counter += 1

       else:
           if ms.rg_cansmi not in rgs_mit_duplikaten:
              rgs_mit_duplikaten.append(ms.rg_cansmi)

           dict_rg_cansmi_titles[ms.rg_cansmi].append(ms.title)
           duplfrei_name = dict_rg_cansmi_duplfrei_title[ms.rg_cansmi]

           Mol_Single.set_duplfrei_name(ms, duplfrei_name)
           Mol_Single.set_duplfrei_mol_pos_nr(ms, duplfrei_name)


    for ms in mol_single_list:
       dupl_list = dict_rg_cansmi_titles[ms.rg_cansmi]
       Mol_Single.set_duplfrei_mols(ms, dupl_list)

    return  rg_duplfrei_liste,  dict_duplfrei_title_2_mcs_pos,  dict_rg_cansmi_duplfrei_title,
dict_rg_cansmi_titles, rgs_mit_duplikaten, check_liste_duplicate
```

```
def fill_mcs_matrices(rg_duplfrei_liste, rgs_mit_duplikaten, mcs_algorithm):
    """
    berechnet von allen mols paarweise den MCS und speichert diesen in mcs_matrix, sowie die
entsprechende NumAtoms in num_atoms_list
    holt sich aus Hauptprogramm: rg_duplfrei_list (enhaelt mols, keine smiles)
    benoetigt Funktion: check_replace_hydrogen()
    Rueckgabe: mcs_matrix, mcs_len_matrix
    """
    n = len(rg_duplfrei_liste)
    mcs_matrix = st.create_leermatrix(n,n)
    mcs_len_matrix = st.create_leermatrix(n,n)
    mcs_len_liste = [] #welche Laengen haben die gefundenen mcs
    all_unique_mcs_liste = [] #enthaelt alle mcs einmal

    for i in range(0,n-1):

        mol_1 = rg_duplfrei_liste[i] #pattern = query
        rg1_cansmi = OECreateCanSmiString(mol_1)

        if rg1_cansmi in rgs_mit_duplikaten:
            #print i, rg1_cansmi
            num_atoms = mol_1.NumAtoms()
            mcs_matrix[i][i] = [rg1_cansmi]
            mcs_len_matrix[i][i] = [num_atoms]
            if rg1_cansmi not in all_unique_mcs_liste:
                all_unique_mcs_liste.append(rg1_cansmi)
            if num_atoms not in mcs_len_liste:
                mcs_len_liste.append(num_atoms)

        for j in range(i+1,n):

            mol_2 = rg_duplfrei_liste[j] #target

            atomexpr = OEExprOpts_DefaultAtoms
            bondexpr = OEExprOpts_DefaultBonds

            mcss = OEMCSSearch(mol_1,atomexpr,bondexpr,mcs_algorithm) #exakt vs. approximate
algorithm
            mcss.SetMCSFunc(OEMCSMaxAtoms())
            mcss.SetMinAtoms(3) #Minimal-Anzahl an Pseudoatomen = 3
            unique = True
            smile_list = [] #lokal
            num_atoms_list = [] #lokal

            for match in mcss.Match(mol_2, unique):
                m = OEGraphMol()
                OESubsetMol(m,match,True)
                mcs_cansmi = OECreateCanSmiString(m)
                mcs_cansmi_checked = check_replace_hydrogen(mcs_cansmi)
                mol_k = OEGraphMol() ##
                OEParseSmiles(mol_k, mcs_cansmi_checked) ##
                mcs_cansmi_checked = OECreateCanSmiString(mol_k) ##
                num_atoms = mol_k.NumAtoms() #neu 251011
                if num_atoms not in mcs_len_liste:
                    mcs_len_liste.append(num_atoms)
                if mcs_cansmi_checked not in smile_list:
                    smile_list.append(mcs_cansmi_checked)
                if num_atoms not in num_atoms_list: ##???eigentlich ueberfluessig! -26-9-12-
                    num_atoms_list.append(num_atoms)
                if mcs_cansmi_checked not in all_unique_mcs_liste:
                    all_unique_mcs_liste.append(mcs_cansmi_checked)


            mcs_matrix[i][j] = smile_list
            mcs_matrix[j][i] = smile_list
            mcs_len_matrix[i][j] = num_atoms_list
            mcs_len_matrix[j][i] = num_atoms_list

    mcs_len_liste.sort()
    print "number of unique MCSs", len(all_unique_mcs_liste)

    return mcs_matrix, mcs_len_matrix, mcs_len_liste, all_unique_mcs_liste


def check_replace_hydrogen(smi):
    """
    ersetzt in smiles noch vorhandene H-Atome, die z.B. bei MCS-Berechnung zustande kommen
```

```
    input: zu ersetzender smile smi
    Rueckgabe: H-freier smile smi
    """
    if "H" in smi:
        smi = smi.replace("H3", "")
        smi = smi.replace("H2", "")
        smi = smi.replace("H4", "")   #neu 251011
        smi = smi.replace("H5", "")
        smi = smi.replace("H6", "")
        smi = smi.replace("H7", "")
        smi = smi.replace("H8", "")
        smi = smi.replace("H9", "")
        if "Hg" in smi:
            smi = smi.replace("Hg", "11")
        if "Hf" in smi:
            smi = smi.replace("Hf", "22")
        smi = smi.replace("H", "")
        if "11" in smi:
            smi = smi.replace("11", "Hg")
        if "22" in smi:
            smi = smi.replace("22", "Hf")

    return smi



def  postprocessing_mcs_matrix_zinc(rg_duplfrei_liste,  mcs_dict_smile_2_counter,  mcs_matrix,
mcs_len_matrix, mcs_len_liste, all_unique_mcs_liste, min_mcs_size):

    n = len(rg_duplfrei_liste)
    new_mcs_matrix = st.create_leermatrix(n,n)
    new_mcs_len_matrix = st.create_leermatrix(n,n)
    new_mcs_len_liste = [] #welche Laengen haben die gefundenen mcs
    new_all_unique_mcs_liste = [] #enthaelt alle mcs einmal

    for i in range(0,n-1):

        for j in range(i+1,n):

            if mcs_len_matrix[i][j] != []:
                len_mcs = mcs_len_matrix[i][j][0]
            else:
                len_mcs = 0
            if len_mcs >= min_mcs_size:
                new_smiles_list = []
                for mcs in mcs_matrix[i][j]:
                    if mcs not in mcs_dict_smile_2_counter:
                        new_smiles_list.append(mcs)
                        if mcs not in new_all_unique_mcs_liste:
                            new_all_unique_mcs_liste.append(mcs)
                            if len_mcs not in new_mcs_len_liste:
                                new_mcs_len_liste.append(len_mcs)
                    else:
                        if  mcs_dict_smile_2_counter[mcs]  <  1000:  ##entspricht  0,1  Prozent
Wahrscheinlichkeit fuer unspezifischen MCS
                            new_smiles_list.append(mcs)
                            if mcs not in new_all_unique_mcs_liste:
                                new_all_unique_mcs_liste.append(mcs)
                                if len_mcs not in new_mcs_len_liste:
                                    new_mcs_len_liste.append(len_mcs)

                new_mcs_matrix[i][j] = new_smiles_list
                new_mcs_matrix[j][i] = new_smiles_list
                if new_smiles_list != []:
                    new_mcs_len_matrix[i][j] = mcs_len_matrix[i][j]
                    new_mcs_len_matrix[j][i] = mcs_len_matrix[j][i]
                else:
                    new_mcs_len_matrix[i][j] = []
                    new_mcs_len_matrix[j][i] = []


            else:
                new_mcs_matrix[i][j] = []
                new_mcs_matrix[j][i] = []
                new_mcs_len_matrix[i][j] = []
                new_mcs_len_matrix[j][i] = []

    mcs_len_liste.sort()
```

```
    print "min_mcs_size", min_mcs_size, "black-list = active"
    print "number of unique MCSs", len(new_all_unique_mcs_liste)

    return new_mcs_matrix, new_mcs_len_matrix, new_mcs_len_liste, new_all_unique_mcs_liste


def    postprocessing_mcs_matrix(rg_duplfrei_liste,    mcs_dict_smile_2_counter,    mcs_matrix,
mcs_len_matrix, mcs_len_liste, all_unique_mcs_liste, min_mcs_size):

    n = len(rg_duplfrei_liste)
    new_mcs_matrix = st.create_leermatrix(n,n)
    new_mcs_len_matrix = st.create_leermatrix(n,n)
    new_mcs_len_liste = mcs_len_liste[:] #welche Laengen haben die gefundenen mcs
    new_all_unique_mcs_liste = [] #enthaelt alle mcs einmal

    for i in range(0,n-1):

        for j in range(i+1,n):

            if mcs_len_matrix[i][j] != []:
                len_mcs = mcs_len_matrix[i][j][0]
            else:
                len_mcs = 0
            if len_mcs >= min_mcs_size:
                for mcs in mcs_matrix[i][j]:
                    if mcs not in new_all_unique_mcs_liste:
                        new_all_unique_mcs_liste.append(mcs)

                    else:
                        if mcs not in new_all_unique_mcs_liste:
                            new_all_unique_mcs_liste.append(mcs)

                new_mcs_matrix[i][j] = mcs_matrix[i][j]
                new_mcs_matrix[j][i] = mcs_matrix[j][i]
                new_mcs_len_matrix[i][j] = mcs_len_matrix[i][j]
                new_mcs_len_matrix[j][i] = mcs_len_matrix[j][i]


            else:
                new_mcs_matrix[i][j] = []
                new_mcs_matrix[j][i] = []
                new_mcs_len_matrix[i][j] = []
                new_mcs_len_matrix[j][i] = []

    for z in range(3,min_mcs_size):
        if z in new_mcs_len_liste:
            new_mcs_len_liste.remove(z)

    print "min_mcs_size", min_mcs_size, "black-list = inactive"
    print "number of unique MCSs", len(new_all_unique_mcs_liste)

    return new_mcs_matrix, new_mcs_len_matrix, new_mcs_len_liste, new_all_unique_mcs_liste


def     pickle_duplfrei_dicts(dict_duplfrei_title_2_mcs_pos,     dict_rg_cansmi_duplfrei_title,
dict_rg_cansmi_titles, rgs_mit_duplikaten, rg_data_name, rg_duplfrei_smiles_liste):
    dateiname = "duplfrei_dicts_and_list_"+rg_data_name+".pkl"
    gesamt_liste_2_pickle = []
    gesamt_liste_2_pickle.append(dict_duplfrei_title_2_mcs_pos)
    gesamt_liste_2_pickle.append(dict_rg_cansmi_duplfrei_title)
    gesamt_liste_2_pickle.append(dict_rg_cansmi_titles)
    gesamt_liste_2_pickle.append(rg_duplfrei_smiles_liste)
    gesamt_liste_2_pickle.append(rgs_mit_duplikaten)

    st.dump_pickle_data(dateiname, gesamt_liste_2_pickle)


def    pickle_mcs_data(mcs_matrix,    mcs_len_matrix,    mcs_len_liste,    all_unique_mcs_liste,
rg_data_name, mcs_size, zinc, algorithm):
    dateiname                                                                                =
"mcs_matrix_data_MCS="+str(mcs_size)+"_blacklist="+str(zinc)+"_Algorithm="+str(algorithm)+"_"+
rg_data_name+".pkl"
    gesamt_liste_2_pickle = []
    gesamt_liste_2_pickle.append(mcs_matrix)
    gesamt_liste_2_pickle.append(mcs_len_matrix)
    gesamt_liste_2_pickle.append(mcs_len_liste)
    gesamt_liste_2_pickle.append(all_unique_mcs_liste)
```

```
    st.dump_pickle_data(dateiname, gesamt_liste_2_pickle)


def check_unique_RG_data_available(rg_data_name):
    start_dir          = os.getcwd()
    all_files          = os.listdir(start_dir)

    dateiname = "duplfrei_dicts_and_list_"+rg_data_name+".pkl"

    all_chunk_folders = [name for name in all_files if dateiname in name]

    if all_chunk_folders != []:
        check = True
    else:
        check = False

    return check


def load_unique_RG_data(rg_data_name):
    dateiname = "duplfrei_dicts_and_list_"+rg_data_name+".pkl"

    gesamt_liste = st.load_pickle_data(dateiname)
    dict_duplfrei_title_2_mcs_pos = gesamt_liste[0]
    dict_rg_cansmi_duplfrei_title = gesamt_liste[1]
    dict_rg_cansmi_titles = gesamt_liste[2]
    rg_duplfrei_smiles_liste = gesamt_liste[3]
    rgs_mit_duplikaten = gesamt_liste[4]

    rg_duplfrei_liste = []
    for rg_smile in rg_duplfrei_smiles_liste:
        rgmol = OEGraphMol()
        OEParseSmiles(rgmol, rg_smile)
        rg_duplfrei_liste.append(rgmol)

    return            dict_duplfrei_title_2_mcs_pos,            dict_rg_cansmi_duplfrei_title,
dict_rg_cansmi_titles, rg_duplfrei_liste, rgs_mit_duplikaten


def check_basic_MCS_data_available(rg_data_name, mcs_size, zinc, algorithm):
    start_dir          = os.getcwd()
    all_files          = os.listdir(start_dir)

    dateiname                                                                                =
"mcs_matrix_data_MCS="+str(mcs_size)+"_blacklist="+str(zinc)+"_Algorithm="+str(algorithm)+"_"+
rg_data_name+".pkl"

    all_chunk_folders = [name for name in all_files if dateiname in name]

    if all_chunk_folders != []:
        check = True
    else:
        check = False

    return check


def load_basic_MCS_data(rg_data_name, mcs_size, zinc, algorithm):
    dateiname                                                                                =
"mcs_matrix_data_MCS="+str(mcs_size)+"_blacklist="+str(zinc)+"_Algorithm="+str(algorithm)+"_"+
rg_data_name+".pkl"

    gesamt_liste = st.load_pickle_data(dateiname)
    mcs_matrix = gesamt_liste[0]
    mcs_len_matrix = gesamt_liste[1]
    mcs_len_liste = gesamt_liste[2]
    all_unique_mcs_liste = gesamt_liste[3]

    return mcs_matrix, mcs_len_matrix, mcs_len_liste, all_unique_mcs_liste


def main():

    usage = "usage: %prog [options], for help use: %prog --help"

    parser = OptionParser(usage)

    start_dir          = os.getcwd()
```

```
    all_files         = os.listdir(start_dir)

    rg_sdf_data               =      [name      for      name      in      all_files      if
'sdwash_sdfilter_duplfrei_fps_added_fp2_rgs.sdf' in name]

    parser.set_defaults(filename=rg_sdf_data[0])
    parser.add_option("-f",    "--file",    dest="filename",    action="store",    type="string",
help="RG-FILENAME.sdf")

    parser.set_defaults(min_atom_size=3)
    parser.add_option("-m",     "--min_mcs_size",     dest="min_atom_size",     action="store",
type="int", help="minimum mcs-size (int)")

    parser.set_defaults(zinc=True)
    parser.add_option("-z",  "--zinc_results",  dest="zinc",  action="store_false",  help="set
blacklist=INACTIVE => KEEP unspecific MCSs")

    parser.set_defaults(mcs_algorithm=True)
    parser.add_option("-a",   "--mcs_algorithm",   dest="mcs_algorithm",   action="store_false",
help="approximate instead of exact mcs algorithm")

    parser.add_option("-v", "--verbose", action="store_true", dest="verbose")
    parser.add_option("-q", "--quiet", action="store_false", dest="verbose")

    (options, args) = parser.parse_args()


    if options.verbose:
        print "reading %s..." % options.filename


    #new: 19-02-13: lade unspezifische MCSs Daten (ZINC Analyse)
    mcs_dict_smile_2_counter                                                                =
st.load_pickle_data("mcs_dict_smile_2_counter_pickeled_gesamt.pkl")

    mol_single_list = []

    #start-zeit
    start, date = st.note_start_time()

    #RG data einlesen
    #rg_data_name =  options.filename.split(".")[0]
    rg_data_name = options.filename.split(".")[0]
    mollist = st.einlesen_der_mols(rg_data_name)

    for mol in mollist:
       new_mol = Mol_Single(mol, "training")
       mol_single_list.append(new_mol)

    min_atom_size = options.min_atom_size
    zinc = options.zinc

    print "min_mcs_size", min_atom_size
    print "specific (black-list)", options.zinc

    #pruefe, ob unique RG liste schon vorhanden fuer diese sdf-Datei
    unique_RG_data_check = check_unique_RG_data_available(rg_data_name)
    if unique_RG_data_check == True:
       dict_duplfrei_title_2_mcs_pos,   dict_rg_cansmi_duplfrei_title,   dict_rg_cansmi_titles,
rg_duplfrei_liste, rgs_mit_duplikaten = load_unique_RG_data(rg_data_name)
    else:
       #pruefe  rgs  auf  duplikate  mittels  kanonischer  SMILES  ->  erstelle  Liste  aller
einzigartigen RGs -> MCS-Berechnung beschleunigen!
       rg_duplfrei_liste,     dict_duplfrei_title_2_mcs_pos,     dict_rg_cansmi_duplfrei_title,
dict_rg_cansmi_titles,          rgs_mit_duplikaten,          rg_duplfrei_smiles_liste          =
check_rg_duplicate(mol_single_list)
       print "Anzahl aller duplfreien rgs", len(rg_duplfrei_liste)

       pickle_duplfrei_dicts(dict_duplfrei_title_2_mcs_pos,     dict_rg_cansmi_duplfrei_title,
dict_rg_cansmi_titles, rgs_mit_duplikaten, rg_data_name, rg_duplfrei_smiles_liste)
       print "duplfrei-data gepickelt"


    #fuer MCS-Berechnung: Wahl des Algorithmus
    if options.mcs_algorithm == True:
       mcs_algorithm = OEMCSType_Exhaustive
       print "exact\n"
    else:
```

```
        mcs_algorithm = OEMCSType_Approximate
        print "approximate\n"


    basic_MCS_data_check = check_basic_MCS_data_available(rg_data_name, 3, False, True)
    if basic_MCS_data_check == True:
        mcs_matrix,      mcs_len_matrix,      mcs_len_liste,      all_unique_mcs_liste      =
load_basic_MCS_data(rg_data_name, 3, False, True)
    else:
        #paarweise MCS berechnen aller einzigartigen RGs
        mcs_matrix,      mcs_len_matrix,      mcs_len_liste,      all_unique_mcs_liste      =
fill_mcs_matrices(rg_duplfrei_liste, rgs_mit_duplikaten, options.mcs_algorithm)

        pickle_mcs_data(mcs_matrix,   mcs_len_matrix,   mcs_len_liste,   all_unique_mcs_liste,
rg_data_name, 3, False, options.mcs_algorithm)
        print "mcs-data gepickelt"

        new_mcs_matrix,   new_mcs_len_matrix,   new_mcs_len_liste,   new_all_unique_mcs_liste   =
postprocessing_mcs_matrix_zinc(rg_duplfrei_liste,     mcs_dict_smile_2_counter,     mcs_matrix,
mcs_len_matrix, mcs_len_liste, all_unique_mcs_liste, 3)

        pickle_mcs_data(new_mcs_matrix,          new_mcs_len_matrix,          new_mcs_len_liste,
new_all_unique_mcs_liste, rg_data_name, 3, True, options.mcs_algorithm)

    if zinc == True:
        new_mcs_matrix,   new_mcs_len_matrix,   new_mcs_len_liste,   new_all_unique_mcs_liste   =
postprocessing_mcs_matrix_zinc(rg_duplfrei_liste,     mcs_dict_smile_2_counter,     mcs_matrix,
mcs_len_matrix, mcs_len_liste, all_unique_mcs_liste, min_atom_size)
    else:
        new_mcs_matrix,   new_mcs_len_matrix,   new_mcs_len_liste,   new_all_unique_mcs_liste   =
postprocessing_mcs_matrix(rg_duplfrei_liste,       mcs_dict_smile_2_counter,       mcs_matrix,
mcs_len_matrix, mcs_len_liste, all_unique_mcs_liste, min_atom_size)

    pickle_mcs_data(new_mcs_matrix,           new_mcs_len_matrix,           new_mcs_len_liste,
new_all_unique_mcs_liste, rg_data_name, min_atom_size, zinc, options.mcs_algorithm)

    print "---finished!!!---"


##================================Beginn                                                    des
Hauptprogrammes=====================================================##


if __name__ == "__main__":
    main()




classes_for_inSARa_3.py

#!/usr/bin/python
#-*- coding: utf-8 -*-
"""
(c) Sabrina Wollenhaupt - eigene Klassen fuer MCS-Berechnung und Netzwerk-Erzeugung
"""

##Modulimport:
from openeye.oechem import *
import cPickle as pickle
import sys, os
import networkx as nx
import math
import random
import time
import numpy as np
import umrechnungen_activity as ua #eigenes Modul
import standard_tasks as st #eigenes Modul
print "\n", "Module erfolgreich importiert", "\n"


class MCS_Node(object):
    def __init__(self, smile, name, mcsmol, atomnum):
        self.smile = smile
        mcsmol = OEGraphMol()
        OEParseSmiles(mcsmol, smile)
        self.mcsmol = mcsmol
```

```
        self.name = name
        self.ebene = atomnum

        self.vorgaenger_node = []

        self.nachfolger_node = []
        self.vorgaenger_nach_mst = []
        self.nachfolger_nach_mst = []
        self.zugehoerige_mols = []
        self.zugehoerige_mols_ms = []
        self.steckenbleibende_mols_ms = []
        self.mean_activity_per_node = None
        self.median_activity_per_node = None
        self.min_activity_per_node = None
        self.max_activity_per_node = None
        self.differenz_min_max_activity_per_node = None
        self.zugehoerige_test_mols_ms = []
        self.steckenbleibende_test_mols_ms = []


        self.vorgaenger = [] #vorgaenger=MCS_Node, der einen kleineren MCS hat als dieser
MCS_Node (dichter an root dran); root hat keinen vorgaenger // enthaelt den namen
        self.nachfolger = [] #nachfolger=MCS_Node, der einen groesseren MCS hat als dieser
MCS_Node
        self.matching_ms = []
        self.unused_ms = []

        self.gini = None
        self.act_class_list = []
        self.winner_class = None

        self.tc_ecfp4 = None
        self.tc_maccs = None
        self.tc_maccsf = None
        self.tc_fp2 = None


    def finde_vorgaenger(self, mcs_nodes_list):
        mcsmol = self.mcsmol #groesserer mcs
        ebene = self.ebene
        stop = 0
        lower_ebene_mcs_nodes_list = []
        for mcsn in mcs_nodes_list:
            if mcsn.ebene < ebene:
                lower_ebene_mcs_nodes_list.append(mcsn)
            if mcsn.ebene == ebene-1:
                ss = OESubSearch(mcsn.smile)
                if ss.SingleMatch(mcsmol):
                    self.vorgaenger.append(mcsn.name)
                    self.vorgaenger_node.append(mcsn)
                    mcsn.nachfolger.append(self.name)
                    mcsn.nachfolger_node.append(self)
                    stop = 1
        if stop == 0:
            while stop == 0:
                ebene = ebene-1
                for mcsn in lower_ebene_mcs_nodes_list:
                    if mcsn.ebene == ebene-1:
                        ss = OESubSearch(mcsn.smile)
                        if ss.SingleMatch(mcsmol):
                            self.vorgaenger.append(mcsn.name)
                            self.vorgaenger_node.append(mcsn)
                            mcsn.nachfolger.append(self.name)
                            mcsn.nachfolger_node.append(self)
                            stop = 1



    def finde_mols(self, rg_duplfrei_liste, mol_single_list):
        """
        bestimmt, welche mols durch welchen MCS repraesentiert werden
        """
        counter = 0
        for rgmol in rg_duplfrei_liste:
            ss = OESubSearch(self.smile) #kleinere
            if ss.SingleMatch(rgmol): #groessere
                for ms in mol_single_list:
                    if ms.duplfrei_mol_pos_nr == counter:
```

```
                    self.zugehoerige_mols.append(ms.title)
                    self.zugehoerige_mols_ms.append(ms)
                    Mol_Single.find_mcs_node(ms, self)
            counter += 1

    def finde_steckenbleibende_mols(self, ms):
        if ms not in self.steckenbleibende_mols_ms:
            self.steckenbleibende_mols_ms.append(ms)

    def finde_steckenbleibende_test_mols(self, ms):
        if ms not in self.steckenbleibende_test_mols_ms:
            self.steckenbleibende_test_mols_ms.append(ms)

    def calculate_mean_median_min_max_activity_per_mcs_node(self):
        activities_list = []
        for steckenbleib_ms in self.steckenbleibende_mols_ms:
            activities_list.append(steckenbleib_ms.p_activity)
        if activities_list != []:
            self.mean_activity_per_node = np.mean(np.array(activities_list))
            self.median_activity_per_node = np.median(np.array(activities_list))
            self.min_activity_per_node = np.min(np.array(activities_list))
            self.max_activity_per_node = np.max(np.array(activities_list))
            self.differenz_min_max_activity_per_node     =     self.max_activity_per_node     -
self.min_activity_per_node

    ###
    def set_nachfolger(self, nachfolger_title):
        if nachfolger_title not in self.nachfolger:
            self.nachfolger.append(nachfolger_title)

    def set_vorgaenger(self, vorgaenger_title):
        if vorgaenger_title not in self.vorgaenger:
            self.vorgaenger.append(vorgaenger_title)

    def remove_vorgaenger(self):
        self.vorgaenger = []

    def remove_nachfolger(self):
        self.nachfolger = []

    def add_ms_to_mcsn(self, ms):
        self.matching_ms.append(ms)
        self.unused_ms.append(ms)

    def del_ms_from_used_list(self, ms):
        self.unused_ms.remove(ms)

    def rename_node(self, new_name):
        self.name = new_name

    def add_gini(self, gini, winner_class):
        self.gini = gini
        self.winner_class = winner_class

    def add_act_class_list(self, mol_class):
        self.act_class_list = mol_class

    def entferne_nachfolger(self):
        for mcsn in self.nachfolger:
            mcsn.vorgaenger.remove(self)
            self.nachfolger.remove(mcsn)
            MCS_Node.entferne_nachfolger(mcsn)

    def    copy_features(self,    vorgaenger_node,    nachfolger_node,    vorgaenger_nach_mst,
nachfolger_nach_mst,    zugehoerige_mols,    zugehoerige_mols_ms,    steckenbleibende_mols_ms,
mean_activity_per_node,            median_activity_per_node,           min_activity_per_node,
max_activity_per_node,     differenz_min_max_activity_per_node,     zugehoerige_test_mols_ms,
steckenbleibende_test_mols_ms,   vorgaenger,   nachfolger,   matching_ms,   unused_ms,   gini,
act_class_list, winner_class):
        self.vorgaenger_node = vorgaenger_node

        self.nachfolger_node = nachfolger_node
        self.vorgaenger_nach_mst = vorgaenger_nach_mst
        self.nachfolger_nach_mst = nachfolger_nach_mst
        self.zugehoerige_mols = zugehoerige_mols
        self.zugehoerige_mols_ms = zugehoerige_mols_ms
        self.steckenbleibende_mols_ms = steckenbleibende_mols_ms
        self.mean_activity_per_node = mean_activity_per_node
```

```
        self.median_activity_per_node = median_activity_per_node
        self.min_activity_per_node = min_activity_per_node
        self.max_activity_per_node = max_activity_per_node
        self.differenz_min_max_activity_per_node = differenz_min_max_activity_per_node
        self.zugehoerige_test_mols_ms = zugehoerige_test_mols_ms
        self.steckenbleibende_test_mols_ms = steckenbleibende_test_mols_ms


        self.vorgaenger = vorgaenger #vorgaenger=MCS_Node, der einen kleineren MCS hat als
dieser MCS_Node (dichter an root dran); root hat keinen vorgaenger // enthaelt den namen
        self.nachfolger = nachfolger #nachfolger=MCS_Node, der einen groesseren MCS hat als
dieser MCS_Node
        self.matching_ms = matching_ms
        self.unused_ms = unused_ms

        self.gini = gini
        self.act_class_list = act_class_list
        self.winner_class = winner_class

    def set_tc(self, mean_tc, fp_type):
        fp_type_list = ["ECFP4", "MACCSF", "MACCS", "FP2_1024"]
        if fp_type == fp_type_list[0]:
            self.tc_ecfp4 = mean_tc
        elif fp_type == fp_type_list[1]:
            self.tc_maccsf = mean_tc
        elif fp_type == fp_type_list[2]:
            self.tc_maccs = mean_tc
        elif fp_type == fp_type_list[3]:
            self.tc_fp2 = mean_tc




class Mol_Single(object):
    mol_pos_counter = 0
    def __init__(self, mol, dataset_status):
        self.title = OEGetSDData(mol, "TITLE")
        self.name = self.title

        self.dataset_status = dataset_status #training=0 oder test=1
        self.original_dataset_status = dataset_status #0=training, 1=test

        try:
            self.status = int(OEGetSDData(mol, "STATUS"))
        except:
            self.status = None

        self.act_class = None

        omol_smile = OEGetSDData(mol, "ISOCANSMI")
        omol = OEGraphMol()
        OEParseSmiles(omol, omol_smile)
        self.mol = omol
        cansmi = OECreateCanSmiString(omol)
        self.cansmi = cansmi
        self.isocansmi = omol_smile

        #rg_smile = OEGetSDData(mol, "RG_SMILE")
        #rgmol = OEGraphMol()
        #OEParseSmiles(rgmol, rg_smile)
        self.rgmol = mol
        rg_cansmi = OECreateCanSmiString(mol)
        self.rg_cansmi = rg_cansmi

        self.activity_nM = OEGetSDData(mol, "ACTIVITY_NM")
        ##umrechnen neues Modul-> Fkt import
        try:
            p_activity = float(OEGetSDData(mol, "P_ACTIVITY")) #kann man, wenn man will in der
sdf-Datei bereitstellen
        except:
            p_activity   =   ua.calculate_pki_from_nM_ki(self.activity_nM)   #kann   aber   auch
berechnet werden
        self.p_activity = p_activity

        try:
            self.activity_status = int(OEGetSDData(mol, "OUTCOME")) #0=schwach, 1=hoch
        except:
```

```
        if self.p_activity >= 7:
            self.activity_status = 1

        else:
            self.activity_status = 0

    self.duplfrei_name = None
    self.duplfrei_mols = []
    self.duplfrei_mol_pos_nr = None

    self.mcs_nodes = []
    self.mcs_nodes_mcsn = []
    self.steckenbleiben_mcs_nodes = []
    self.steckenbleiben_mcs_nodes_mcsn = []
    self.mol_pos_nr = Mol_Single.mol_pos_counter
    Mol_Single.mol_pos_counter += 1

    self.represented = 0 #0 = nicht mcs_node zugeordnet, 1 = repraesentiert
    self.mol_node_names = []
    self.rg_fp = None
    self.maccs_fp = None

    #print self.title, self.cansmi, self.rg_cansmi, self.activity_nM, "\n"


def set_mols_attribute_to_none(self):
    """
    sonst ist kein Picklen der MS-Objekte moeglich
    """
    self.rgmol = None
    self.mol = None

def add_mol_attribute(self, cansmi):
    mol = OEGraphMol()
    OEParseSmiles(mol, cansmi)
    self.mol = mol

def add_rg_mol_attribute(self, rg_cansmi):
    rgmol = OEGraphMol()
    OEParseSmiles(rgmol, rg_cansmi)
    self.rgmol = rgmol

def init_missing_attributes(self, mol_pos_counter):
    self.duplfrei_name = None
    self.duplfrei_mols = []
    self.duplfrei_mol_pos_nr = None

    self.mcs_nodes = []
    self.mcs_nodes_mcsn = []
    self.steckenbleiben_mcs_nodes = []
    self.steckenbleiben_mcs_nodes_mcsn = []
    self.mol_pos_nr = mol_pos_counter

    self.represented = 0 #0 = nicht mcs_node zugeordnet, 1 = repraesentiert
    self.mol_node_names = []

def change_current_dataset_status(self, status):
    self.dataset_status = status


def find_mcs_node(self, mcsn):
    self.mcs_nodes.append(mcsn.name)
    self.mcs_nodes_mcsn.append(mcsn)
    if self.represented != 1:
        Mol_Single.change_represented_status(self)

def find_steckenbleiben_mcs_nodes(self, steckenbleib_nodes):
    self.steckenbleiben_mcs_nodes = steckenbleib_nodes[:]
    for mcsn in steckenbleib_nodes:
          MCS_Node.finde_steckenbleibende_mols(mcsn, self)

def search_rg_duplicates():
    pass

def get_mol_nr():
    pass
```

```
    ### Funktion duplfreie Mols
    def set_duplfrei_name(self, duplfrei_name):
        self.duplfrei_name = duplfrei_name

    def set_duplfrei_mol_pos_nr(self, duplfrei_name):
        splitted_name = duplfrei_name.split("_")
        pos_nr = splitted_name[1]
        self.duplfrei_mol_pos_nr = int(pos_nr)


    def set_duplfrei_mols(self, duplfrei_list):
        for title in duplfrei_list:
            if title not in self.duplfrei_mols:
                self.duplfrei_mols.append(title)
    ###



    def change_represented_status(self):
        self.represented = 1

    def add_mol_node_name(self, node_name):
        self.mol_node_names.append(node_name)

    ###

    def set_activity_class(self):
        if self.p_activity >=7:
            self.act_class = "H"
        elif self.p_activity < 7:
            self.act_class = "L"
        else:
            print "selectivity-class-error"


    def del_mcs_nodes(self):
        self.mcs_nodes = []
        self.mcs_nodes_mcsn = []
```

27.3. Netzwerk-Erzeugung

```
inSARa_network_generation_mit_Kommandozeilenfunktion.py

#!/usr/bin/python
#-*- coding: utf-8 -*-

"""
(c) Sabrina Wollenhaupt - inSARa for SAR-Analysis (Network-Generation)
Version 0.02 - 10-27-2013
"""


##Modulimport: am Ende mal sehen, welche man wirklich braucht
from openeye.oechem import *
import cPickle as pickle
import sys, os
import networkx as nx
import math #math.log(x,2)
import random
import time
import numpy as np
import standard_tasks as st #eigenes Modul (Standardaufgaben)
from classes_for_insara_3 import *

from optparse import OptionParser ##Optionen fuer Kommandozeilen-Funktion

print "\n", "Module erfolgreich importiert", "\n"



def check_rg_duplicate(mol_single_list):
    """
    sucht nach Duplikaten bei den RGs und erstellt eine rg_duplfrei_liste (um Effizienz der
MCS-Rechnungen zu steigern)
```

```
    """
    check_liste_duplicate = []
    dict_rg_cansmi_titles = {}
    dict_rg_cansmi_duplfrei_title = {}
    duplfrei_name_counter = 0

    rg_duplfrei_liste = [] #enthaelt rgmols (ohne Duplikate)

    for ms in mol_single_list:
        if ms.rg_cansmi not in check_liste_duplicate:
            check_liste_duplicate.append(ms.rg_cansmi)
            dict_rg_cansmi_titles[ms.rg_cansmi] = [ms.title]

            duplfrei_name = "duplfrei_"+str(duplfrei_name_counter)
            Mol_Single.set_duplfrei_name(ms, duplfrei_name)
            Mol_Single.set_duplfrei_mol_pos_nr(ms, duplfrei_name)

            rgmol = OEGraphMol()
            OEParseSmiles(rgmol, ms.rg_cansmi)
            rg_duplfrei_liste.append(rgmol)

            dict_rg_cansmi_duplfrei_title[ms.rg_cansmi] = duplfrei_name
            duplfrei_name_counter += 1

        else:
            dict_rg_cansmi_titles[ms.rg_cansmi].append(ms.title)
            duplfrei_name = dict_rg_cansmi_duplfrei_title[ms.rg_cansmi]

            Mol_Single.set_duplfrei_name(ms, duplfrei_name)
            Mol_Single.set_duplfrei_mol_pos_nr(ms, duplfrei_name)


    for ms in mol_single_list:
        dupl_list = dict_rg_cansmi_titles[ms.rg_cansmi]
        Mol_Single.set_duplfrei_mols(ms, dupl_list)

    return rg_duplfrei_liste


def find_mcs_without_vorgaenger(all_unique_mcs_liste):
    """
    finde alle mcs ohne kleineren mcs (= vorgaenger)
    """
    #erstelle mcs_nodes_list (alle unique mcs)
    mcs_node_counter = 0
    mcs_nodes_list = [] #enthaelt MCS_Node-Objekte
    max_mcs_size = 0

    for mcs_smile in all_unique_mcs_liste:
        mcsmol = OEGraphMol()
        OEParseSmiles(mcsmol, mcs_smile)
        num_atoms_smile = mcsmol.NumAtoms()
        if num_atoms_smile > max_mcs_size:
            max_mcs_size = num_atoms_smile


        mcs_node_name = "mcs_node_"+str(mcs_node_counter)
        new_mcs_node = MCS_Node(mcs_smile, mcs_node_name, mcsmol, num_atoms_smile)
        mcs_nodes_list.append(new_mcs_node)
        mcs_node_counter += 1

    start_ebene = 3


    for i in range(start_ebene, max_mcs_size):
        for mcsn in mcs_nodes_list:
            if mcsn.nachfolger == []:
                if mcsn.ebene == i:
                    follow_ebene = i+1
                    for j in range(i+1, max_mcs_size):
                        if mcsn.nachfolger == []:
                            for mcsn_follow in mcs_nodes_list:
                                if mcsn_follow.ebene == follow_ebene:
                                    match         =         mache_substruktur_suche(mcsn.smile,
mcsn_follow.smile)
                                    if match == 1:
                                        MCS_Node.set_nachfolger(mcsn, mcsn_follow)
                                        MCS_Node.set_vorgaenger(mcsn_follow, mcsn)
```

```
                        if mcsn.nachfolger == []:
                            if follow_ebene+1 <= max_mcs_size:
                                follow_ebene += 1
                    else:
                        break


    mcsn_ohne_vorgaenger_max_size = 0
    mcsn_ohne_vorgaenger = []
    for mcsn in mcs_nodes_list:

        if mcsn.vorgaenger == []:

            mcsn_ohne_vorgaenger.append(mcsn)
            if mcsn.ebene > mcsn_ohne_vorgaenger_max_size:
                mcsn_ohne_vorgaenger_max_size = mcsn.ebene


    k_list = [k for k in range(start_ebene+1, mcsn_ohne_vorgaenger_max_size+1)]
    k_list.sort(reverse=True)
    for k in k_list:

        for mcsn in mcsn_ohne_vorgaenger:

            if mcsn.vorgaenger == []:
                if mcsn.ebene == k:
                    before_ebene = k-1
                    l_list = [l for l in range(start_ebene+1, k)]
                    l_list.sort(reverse=True)
                    for l in l_list:
                        if mcsn.vorgaenger == []:
                            for mcsn_before in mcs_nodes_list:
                                if mcsn_before.ebene == before_ebene:
                                    match     =     mache_substruktur_suche(mcsn_before.smile,
mcsn.smile)
                                    if match == 1:
                                        #print "yes"
                                        MCS_Node.set_vorgaenger(mcsn, mcsn_before)
                                        MCS_Node.set_nachfolger(mcsn_before, mcsn)
                            if mcsn.vorgaenger == []:
                                if before_ebene-1 >= start_ebene:
                                    before_ebene -= 1
                        else:
                            break

    possible_root_mcs_list = []
    for mcsn in mcs_nodes_list:
        if mcsn.vorgaenger == []:
            possible_root_mcs_list.append(mcsn)


    return possible_root_mcs_list, mcs_nodes_list


def mache_substruktur_suche(kleinere_struktur_smile, groessere_struktur_smile):
    match = 0
    ss = OESubSearch(kleinere_struktur_smile)
    groessere_struktur_mol = OEGraphMol()
    OEParseSmiles(groessere_struktur_mol, groessere_struktur_smile)
    if ss.SingleMatch(groessere_struktur_mol):
        match = 1

    return match


def   search_mols_belonging_to_possible_root_mcss(rg_duplfrei_liste,   possible_root_mcs_list,
mol_single_list):
    """
    sucht jeweils die mols, die zu den ensprechenden moeglichen root-mcs gehoeren
    """
    for mcsn in possible_root_mcs_list: #zuordnung mcs:molpos.
        ss = OESubSearch(mcsn.smile)
        mol_position = 0
        for duplfrei_rg_mol in rg_duplfrei_liste:
            if ss.SingleMatch(duplfrei_rg_mol):
                for ms in mol_single_list:
                    if ms.duplfrei_mol_pos_nr == mol_position:
                        MCS_Node.add_ms_to_mcsn(mcsn, ms)
```

```
            mol_position += 1


def find_first_root_node(possible_root_mcs_list):
    """
    sucht den ersten mcs-root-node
    """
    final_root_nodes_list = []
    used_mols_list = []
    current_winning_mcsn = []
    current_winning_anzahl = 0
    for mcsn in possible_root_mcs_list:
       if len(mcsn.matching_ms) > current_winning_anzahl:
           current_winning_anzahl = len(mcsn.matching_ms)
    for mcsn in possible_root_mcs_list:
       if len(mcsn.matching_ms) == current_winning_anzahl:
           current_winning_mcsn.append(mcsn)

    print "\nfirst root node found"
    for mcsn in current_winning_mcsn:
       print mcsn.smile, len(mcsn.matching_ms)
    #hier fehlt der Fall des Gleichstandes...weniger MCS???
    winning_mcsn = current_winning_mcsn[0]
    final_root_nodes_list.append(winning_mcsn)

    for mcsn_possible in possible_root_mcs_list:
       for ms in winning_mcsn.matching_ms:
           if ms not in used_mols_list:
              used_mols_list.append(ms)
           if ms in mcsn_possible.unused_ms:
              MCS_Node.del_ms_from_used_list(mcsn_possible, ms)

    return used_mols_list, final_root_nodes_list


def determine_if_stopp_kriterium_fullfilled(mol_single_list, used_mols_list, abbruch_factor):
    """
    bestimmen, ob genug root nodes gefunden, um Netzwerk aufzubauen
    """
    stoppe_knoten_sammeln = 0
    number_used_mols = float(len(used_mols_list))
    number_total_mols = float(len(mol_single_list))
    number_unused_mols = number_total_mols - number_used_mols

    output_data = open("doku_possible_root_nodes.txt", "a")

    print "Anzahl unbenutzte mols", number_unused_mols

    unrepresented_anteil = float(number_unused_mols/number_total_mols)
    print "unrepresented_anteil", unrepresented_anteil*100, "Prozent"

    print >> output_data, "Anzahl unbenutzte mols", number_unused_mols
    print >> output_data, "unrepresented_anteil", unrepresented_anteil*100, "Prozent"
    output_data.close()

    if unrepresented_anteil <= abbruch_factor:
       stoppe_knoten_sammeln = 1

    return stoppe_knoten_sammeln


def find_additional_root_nodes(possible_root_mcs_list, used_mols_list, final_root_nodes_list):
    """
    sucht weitere mcs-root-nodes
    """
    print "\n"
    for mcsn in possible_root_mcs_list:
       print mcsn.smile, len(mcsn.unused_ms), len(mcsn.matching_ms)

    current_winning_mcsn = []
    current_winning_anzahl = 0
    for mcsn in possible_root_mcs_list:
       if len(mcsn.unused_ms) > current_winning_anzahl:
           current_winning_anzahl = len(mcsn.unused_ms)
    for mcsn in possible_root_mcs_list:
       if current_winning_anzahl > 0:
           if len(mcsn.unused_ms) == current_winning_anzahl:
              current_winning_mcsn.append(mcsn)
```

```
    output_data = open("doku_possible_root_nodes.txt", "a")
    if len(current_winning_mcsn) > 0:
        print "\nadditional root node found"
        print >> output_data, "\nadditional root node found"
        for mcsn in current_winning_mcsn:
            print mcsn.smile, len(mcsn.unused_ms), len(mcsn.matching_ms)
            print >> output_data, mcsn.smile, len(mcsn.unused_ms), len(mcsn.matching_ms)
    else:
        print "\nno root node providing new molecules found -> stopp root node selection"
        print >> output_data, "\nno root node providing new molecules found -> stopp root node
selection"

    output_data.close()

    if len(current_winning_mcsn) > 0:

        check = 0 #= nicht stoppen
        if len(current_winning_mcsn) == 1:
            winning_mcsn = current_winning_mcsn[0]
            final_root_nodes_list.append(winning_mcsn)

            for mcsn_possible in possible_root_mcs_list:
                for ms in winning_mcsn.matching_ms:
                    if ms not in used_mols_list:
                        used_mols_list.append(ms)
                    if ms in mcsn_possible.unused_ms:
                        MCS_Node.del_ms_from_used_list(mcsn_possible, ms)
        else:
            tie_breaking_list = []
            for mcsn in current_winning_mcsn:
                new_entry = [(len(mcsn.matching_ms)-len(mcsn.unused_ms)), mcsn]
                tie_breaking_list.append(new_entry)
            tie_breaking_list.sort()
            min_num = tie_breaking_list[0][0]
            new_winning_list = []
            for diff_num, mcsn in tie_breaking_list:
                if diff_num == min_num:
                    new_winning_list.append(mcsn)
            output_data = open("doku_possible_root_nodes.txt", "a")

            print >> output_data, "\ntie breaking"
            for mcsn in new_winning_list:
                print >> output_data, mcsn.smile, len(mcsn.unused_ms), len(mcsn.matching_ms)
            output_data.close()

            winning_mcsn = new_winning_list[0]
            final_root_nodes_list.append(winning_mcsn)

            for mcsn_possible in possible_root_mcs_list:
                for ms in winning_mcsn.matching_ms:
                    if ms not in used_mols_list:
                        used_mols_list.append(ms)
                    if ms in mcsn_possible.unused_ms:
                        MCS_Node.del_ms_from_used_list(mcsn_possible, ms)
    else:
        check = 1 #=stopp rode node selection

    return used_mols_list, final_root_nodes_list, check


def suche_weitere_nachfolger(follow_mcsn, all_mcs):
    for mcsn in follow_mcsn.nachfolger:
        if mcsn not in all_mcs:
            all_mcs.append(mcsn)
            all_mcs = suche_weitere_nachfolger(mcsn, all_mcs)
    return all_mcs


def find_mcs_relationships(one_root_all_mcs_liste):
    """
    finde alle mcs ohne kleineren mcs (= vorgaenger)
    """
    #erstelle mcs_nodes_list (alle unique mcs)
    max_mcs_size = 0
    min_mcs_size = 100

    for mcsn in one_root_all_mcs_liste:
```

```
        mcsmol = OEGraphMol()
        OEParseSmiles(mcsmol, mcsn.smile)
        num_atoms_smile = mcsmol.NumAtoms()
        if num_atoms_smile > max_mcs_size:
            max_mcs_size = num_atoms_smile
        if num_atoms_smile < min_mcs_size:
            min_mcs_size = num_atoms_smile



    for i in range(min_mcs_size, max_mcs_size):
        for mcsn in one_root_all_mcs_liste:
            if mcsn.nachfolger == []:
                if mcsn.ebene == i:
                    follow_ebene = i+1
                    for j in range(i+1, max_mcs_size): #warum???-> weil unten follow_ebene
hochgesetzt wird...
                        if mcsn.nachfolger == []:
                            for mcsn_follow in one_root_all_mcs_liste:
                                if mcsn_follow.ebene == follow_ebene:
                                    match = mache_substruktur_suche(mcsn.smile,
mcsn_follow.smile)
                                    if match == 1:
                                        MCS_Node.set_nachfolger(mcsn, mcsn_follow)
                                        MCS_Node.set_vorgaenger(mcsn_follow, mcsn)
                            if mcsn.nachfolger == []:
                                if follow_ebene+1 <= max_mcs_size:
                                    follow_ebene += 1
                        else:
                            break

    mcsn_ohne_vorgaenger_max_size = 0
    mcsn_ohne_vorgaenger = []
    for mcsn in one_root_all_mcs_liste:
        if mcsn.vorgaenger == []:
            mcsn_ohne_vorgaenger.append(mcsn)

            if mcsn.ebene > mcsn_ohne_vorgaenger_max_size:
                mcsn_ohne_vorgaenger_max_size = mcsn.ebene

    k_list = [k for k in range(min_mcs_size, mcsn_ohne_vorgaenger_max_size+1)]
    k_list.sort(reverse=True)
    for k in k_list:
        for mcsn in mcsn_ohne_vorgaenger:
            if mcsn.vorgaenger == []:
                if mcsn.ebene == k:
                    before_ebene = k-1
                    l_list = [l for l in range(min_mcs_size, k)]
                    l_list.sort(reverse=True)
                    for l in l_list:
                        if mcsn.vorgaenger == []:
                            for mcsn_before in one_root_all_mcs_liste:
                                if mcsn_before.ebene == before_ebene:
                                    match = mache_substruktur_suche(mcsn_before.smile,
mcsn.smile)
                                    if match == 1:
                                        MCS_Node.set_vorgaenger(mcsn, mcsn_before)
                                        MCS_Node.set_nachfolger(mcsn_before, mcsn)
                            if mcsn.vorgaenger == []:
                                if before_ebene-1 >= min_mcs_size:
                                    before_ebene -= 1
                        else:
                            break

    possible_root_mcs_list = []
    for mcsn in one_root_all_mcs_liste:
        if mcsn.vorgaenger == []:
            possible_root_mcs_list.append(mcsn)

    if len(possible_root_mcs_list) > 1:
        for mcsn in possible_root_mcs_list:
            print mcsn.smile, mcsn.zugehoerige_mols
        print "error: MCS-Hierarchie nicht richtig established! Bitte ueberpruefen!"
        sys.exit(2)



def build_mcs_network(network_islands_list, final_root_nodes_list, G):
```

```
    root_counter = 0
    for root in final_root_nodes_list:
        all_island_mcs = network_islands_list[root_counter]
        #alle MCS_Knoten erstellen
        for mcsn in all_island_mcs:
            if mcsn not in G.nodes():
                G.add_node(mcsn)

        #MCS_Knoten verbinden ueber nachfolger
        for created_node in G.nodes():
            if created_node.nachfolger != []:
                for nachfolger_mcsn in created_node.nachfolger:
                    #for mcsn in mcs_nodes_list:
                        #if mcsn.name == nachfolger_name:
                            if G.has_edge(created_node, nachfolger_mcsn) == False:
                                G.add_edge(created_node, nachfolger_mcsn)
        root_counter += 1

    print "\nnumber_nodes", G.number_of_nodes()
    print "number_edges", G.number_of_edges()


    return G


def add_all_mols_2_mcs_network(G):
    #mol_nodes erstellen und mit MCS_Knoten verbinden
    for mcs_node in G.nodes():
        for ms in mcs_node.zugehoerige_mols_ms:

            G.add_node(ms)
            G.add_edge(mcs_node, ms)


    print G.number_of_nodes()
    print G.number_of_edges()
    return G


def find_minimum_spanning_tree(G):
    test_mol_nr = 10**6 #variabler!
    for edge in G.edges():
        edge_0_mols = len(edge[0].zugehoerige_mols)
        edge_1_mols = len(edge[1].zugehoerige_mols)
        if edge_0_mols > edge_1_mols or edge_0_mols == edge_1_mols:
            edge_mol_attribute = test_mol_nr - edge_0_mols
        else:
            edge_mol_attribute = test_mol_nr - edge_1_mols
        G[edge[0]][edge[1]]['mol_weight'] = edge_mol_attribute

    print "\n\n"
    MST = nx.minimum_spanning_tree(G,weight='mol_weight')

    print "Kantenzahl G", len(G.edges()), "Kantenzahl MST", len(MST.edges())

    return MST


def export_network_2_cytoscape_MST(Graph, dateiname_network_2_export):
    mol_node_counter = 0
    G_export = nx.Graph()
    for node in Graph.nodes():
        if node.name.split("_")[0] == "mcs" or node.name.split("_")[0] == "root":
            G_export.add_node(node.name)
        #jedes mols vervielfachen und verbinden
        else:
            mol_duplicate_counter = 0
            for nbor in Graph.neighbors(node):
                node_name = "mol_"+str(mol_node_counter)+"_"+str(mol_duplicate_counter)
                Mol_Single.add_mol_node_name(node, node_name)
                G_export.add_node(node_name)
                G_export.add_edge(node_name,nbor.name)
                mol_duplicate_counter += 1
            mol_node_counter += 1

    for edge1, edge2 in Graph.edges():
        if edge1.name.split("_")[0] == "mcs" or edge1.name.split("_")[0] == "root":
```

```
            if edge2.name.split("_")[0] == "mcs" or edge2.name.split("_")[0] == "root":
                G_export.add_edge(edge1.name, edge2.name)


    print "Anzahl export nodes", G_export.number_of_nodes(), "Anzahl export edges",
G_export.number_of_edges()

    nx.write_gml(G_export, dateiname_network_2_export+".gml") #wichtig node_ids duerfen nur
als strings, nicht als int abgespeichert werden, sonst kein Import in Cytoscape moeglich

    return G_export


def create_edgelist_of_mol(ms):
    edge_list_of_mol = []
    edge2 = ms
    for mcsn in ms.mcs_nodes_mcsn:
        edge1 = mcsn
        new_edge = (edge1, edge2)
        edge_list_of_mol.append(new_edge)

    return edge_list_of_mol


def copy_of_Graph(Graph):
    """
    wenn Graph.copy() nicht fkt, damit man trotzdem eine Kopie eines komplexen Graphen
erzeugen kann
    """
    copy_of_Graph = nx.Graph()
    for node in Graph.nodes():
        copy_of_Graph.add_node(node)
    for edge1, edge2 in Graph.edges():
        copy_of_Graph.add_edge(edge1, edge2)
    return copy_of_Graph


def delete_edges_for_steckenbleiben_mols(Graph, mol_single_list):
    """
    entfernt die Kanten zu den MCS, wo das Mol nicht steckenbleibt
    """
    print "Anzahl nodes G:", Graph.number_of_nodes(), "Anzahl edges G:",
Graph.number_of_edges()

    steckenbleib_Graph = copy_of_Graph(Graph) #austricksen: Graph.copy() fkt in diesem Fall
nicht, da zu viele Iterationen (vermutlich zu viele Infos hinterlegt)
    for ms in mol_single_list:
        edge2 = ms
        edges_2_remove = create_edgelist_of_mol(ms)
        for mcsn in ms.steckenbleiben_mcs_nodes:
            edge1 = mcsn
            steckenbleib_edge = (edge1, edge2)
            if steckenbleib_edge in edges_2_remove:
                edges_2_remove.remove(steckenbleib_edge)
        for edge_2_remove1, edge_2_remove2 in edges_2_remove:
            steckenbleib_Graph.remove_edge(edge_2_remove1, edge_2_remove2)

    print "Anzahl steckenbleib G nodes:",steckenbleib_Graph.number_of_nodes(), " Anzahl
steckenbleib G edges:", steckenbleib_Graph.number_of_edges()

    return steckenbleib_Graph



def export_attributes_2_cytoscape(Graph, split, date, rg_data_name):
    dict_molname_2_attributes = {}
    dict_mcs_node_2_attributes = {}

    for node in Graph.nodes():
        if node.name.split("_")[0] != "root" and node.name.split("_")[0] != "mcs":
            for node_name in node.mol_node_names:
                if node_name not in dict_molname_2_attributes:
                    dict_molname_2_attributes[node_name] = [node.p_activity, node.name,
node.cansmi, node.rg_cansmi]
        else:
            if node.name not in dict_mcs_node_2_attributes:
                dict_mcs_node_2_attributes[node.name] = [node.smile, node.gini,
node.winner_class, node.ebene]
```

```
    attribute_ausgabe_datei_1                                                    =
open("attribute_ausgabe_mols_steck_network_"+rg_data_name+"_"+str(date)+".txt", "w")


    print                              >>                                attribute_ausgabe_datei_1,
"Title"+"\t"+"p_Activity"+"\t"+"Mol_Name"+"\t"+"O_SMILE"+"\t"+"RG_SMILE"
    for k,v in dict_molname_2_attributes.iteritems():
       print >> attribute_ausgabe_datei_1, k+"\t"+str(v[0])+"\t"+v[1]+"\t"+v[2]+"\t"+v[3]

    attribute_ausgabe_datei_1.close()


    attribute_ausgabe_datei_2                                                    =
open("attribute_ausgabe_mcsnodes_steck_network_"+rg_data_name+"_"+str(date)+".txt", "w")


    print                              >>                                attribute_ausgabe_datei_2,
"Title"+"\t"+"MCS_SMILE"+"\t"+"gini_index"+"\t"+"act_class"+"\t"+"ebene"
    for k,v in dict_mcs_node_2_attributes.iteritems():
       print                            >>                                attribute_ausgabe_datei_2,
k+"\t"+v[0]+"\t"+str(v[1])+"\t"+str(v[2])+"\t"+str(v[3])

    attribute_ausgabe_datei_2.close()


def determine_steckenbleiben_nodes(mol_single_list):
    for ms in mol_single_list:
       possible_steckenbleiben_nodes = ms.mcs_nodes_mcsn[:]
       for possible_steckenbleib_mcsn in ms.mcs_nodes_mcsn: #hier sind die ganzen MCSN drin
           for vorgaenger in possible_steckenbleib_mcsn.vorgaenger: #vorgaenger hat kleineren
MCS
               if vorgaenger in possible_steckenbleiben_nodes:
                   possible_steckenbleiben_nodes.remove(vorgaenger)
       Mol_Single.find_steckenbleiben_mcs_nodes(ms, possible_steckenbleiben_nodes)


def aktualisiere_vorgaenger_nachfolger_beziehung(G):
    for mcsn in G.nodes():
       #print mcsn.ebene
       MCS_Node.remove_vorgaenger(mcsn)
       MCS_Node.remove_nachfolger(mcsn)

       for nbor_mcsn in G.neighbors(mcsn):
           #print nbor_mcsn.ebene
           if nbor_mcsn.ebene > mcsn.ebene:
              MCS_Node.set_nachfolger(mcsn, nbor_mcsn)
              MCS_Node.set_vorgaenger(nbor_mcsn, mcsn)
           elif nbor_mcsn.ebene < mcsn.ebene:
              MCS_Node.set_nachfolger(nbor_mcsn, mcsn)
              MCS_Node.set_vorgaenger(mcsn, nbor_mcsn)
           else:
              print "ERROR: gleiche EBENE!!!"
              sys.exit(5)



def unpickle_matrix_data_and_duplfrei_data(rg_data_name, zinc, mcs_size):

    start_dir         = os.getcwd()
    all_files         = os.listdir(start_dir)

    print "min_mcs_size=", mcs_size
    print "blacklist=", zinc

    dateiname_1               =       [name       for       name       in       all_files       if
"mcs_matrix_data_MCS="+str(mcs_size)+"_blacklist="+str(zinc)+"_"       in       name       and
"_"+rg_data_name+".pkl" in name]

    gesamt_liste_1 = st.load_pickle_data(dateiname_1[0])
    mcs_matrix = gesamt_liste_1[0]
    mcs_len_matrix = gesamt_liste_1[1]
    mcs_len_liste = gesamt_liste_1[2]
    all_unique_mcs_liste = gesamt_liste_1[3]

    dateiname_2 = "duplfrei_dicts_and_list_"+rg_data_name+".pkl"
    gesamt_liste_2 = st.load_pickle_data(dateiname_2)
```

```
    dict_duplfrei_title_2_mcs_pos = gesamt_liste_2[0]
    dict_rg_cansmi_duplfrei_title = gesamt_liste_2[1]
    dict_rg_cansmi_titles = gesamt_liste_2[2]
    rg_duplfrei_smiles_liste = gesamt_liste_2[3]

    rg_duplfrei_liste = []
    for rg_smile in rg_duplfrei_smiles_liste:
        rgmol = OEGraphMol()
        OEParseSmiles(rgmol, rg_smile)
        rg_duplfrei_liste.append(rgmol)

    return  rg_duplfrei_liste,  dict_duplfrei_title_2_mcs_pos,  dict_rg_cansmi_duplfrei_title,
dict_rg_cansmi_titles, mcs_matrix, mcs_len_matrix, mcs_len_liste, all_unique_mcs_liste


def          determine_unique_training_mcs(mcs_matrix,          dict_duplfrei_title_2_mcs_pos,
dict_rg_cansmi_duplfrei_title, training_mol_single_list, rg_duplfrei_liste):
    all_unique_mcs_liste = []
    n = len(training_mol_single_list)
    for i in range(0,n-1):
        for j in range(i+1,n):
                ms1 = training_mol_single_list[i]
                ms2 = training_mol_single_list[j]
                pos_1                                                                          =
dict_duplfrei_title_2_mcs_pos[dict_rg_cansmi_duplfrei_title[ms1.rg_cansmi]]
                pos_2                                                                          =
dict_duplfrei_title_2_mcs_pos[dict_rg_cansmi_duplfrei_title[ms2.rg_cansmi]]
                mcs_list = mcs_matrix[pos_1][pos_2]

                if pos_1 != pos_2:
                    #print mcs_list
                    if type(mcs_list) != []:
                        for mcs in mcs_list:
                            #print mcs
                            if mcs not in all_unique_mcs_liste:
                                  all_unique_mcs_liste.append(mcs)

    return all_unique_mcs_liste



##=========================Initialisieren           von           globalen           Variablen
etc.=========================================##

mol_single_list = [] #enthaelt Mol_Single-Objekte (training)
stopp = 0
k=3


##================================Beginn                                                      des
Hauptprogrammes=====================================================##

def usage():
    print "USAGE: "+sys.argv[0]+" <dateiname sdf-datei>"
    print "sys.argv[1]: ohne Endung .sdf / muss (RG-)mols enthalten + sdf-tags\n"
    sys.exit(1)



if __name__=='__main__':

        #if len(sys.argv) != 3:
            #usage()

        #start-zeit
        start, date = st.note_start_time()


        usage = "usage: %prog [options], for help use: %prog --help"

        parser = OptionParser(usage)

        start_dir             = os.getcwd()
        all_files             = os.listdir(start_dir)

        rg_sdf_data           =      [name      for      name      in      all_files      if
'sdwash_sdfilter_duplfrei_fps_added_fp2_rgs.sdf' in name]
```

```
        parser.set_defaults(filename=rg_sdf_data[0])
        parser.add_option("-f",   "--file",   dest="filename",   action="store",   type="string",
help="RG-FILENAME.sdf")

        parser.set_defaults(min_atom_size=3)
        parser.add_option("-m",    "--min_mcs_size",    dest="min_atom_size",    action="store",
type="int", help="minimum mcs-size (int)")

        parser.set_defaults(zinc=True)
        parser.add_option("-z", "--zinc_results", dest="zinc", action="store_false", help="set
blacklist=INACTIVE => KEEP unspecific MCSs")


        parser.set_defaults(abbruch_factor=0.02)
        parser.add_option("-s", "--stopp", dest="abbruch_factor", action="store", type='float',
help="termination criterion root-node selection, portion of unrepresented mols (float), e.g.
0.01 (= 1%)")

        parser.set_defaults(number_roots=100)
        parser.add_option("-r", "--num_roots", dest="number_roots", action="store", type='int',
help="max number of root-nodes -> stopp selection (int)")


        parser.add_option("-v", "--verbose", action="store_true", dest="verbose")
        parser.add_option("-q", "--quiet", action="store_false", dest="verbose")

        (options, args) = parser.parse_args()


        abbruch_factor = options.abbruch_factor

        loop_counter = options.number_roots

        min_atom_size = options.min_atom_size
        zinc = options.zinc

        rg_data_name = options.filename.split('.')[0]

        #einlesen der mols (-> mollist enthaelt alle rgmols (training))
        mollist = st.einlesen_der_mols(rg_data_name)

        for mol in mollist:
            new_mol = Mol_Single(mol, "training")
            mol_single_list.append(new_mol)


        date                                                                                    =
str(abbruch_factor*100)+"%_MCS="+str(min_atom_size)+"_blacklist="+str(zinc)+"_"+date


        #load pickled mcs and rg data
        rg_duplfrei_liste,     dict_duplfrei_title_2_mcs_pos,     dict_rg_cansmi_duplfrei_title,
dict_rg_cansmi_titles,   mcs_matrix,   mcs_len_matrix,   mcs_len_liste,   all_unique_mcs_liste   =
unpickle_matrix_data_and_duplfrei_data(rg_data_name, zinc, min_atom_size)


        for ms in mol_single_list:
            duplfrei_name = dict_rg_cansmi_duplfrei_title[ms.rg_cansmi]
            Mol_Single.set_duplfrei_name(ms, duplfrei_name)
            Mol_Single.set_duplfrei_mol_pos_nr(ms, duplfrei_name)
            dupl_list = dict_rg_cansmi_titles[ms.rg_cansmi]
            Mol_Single.set_duplfrei_mols(ms, dupl_list)


        #neues verzeichnis erstellen und reinwechseln
        st.create_new_dir_and_change_dir(date)

        network_islands_list = []

        #duplikate_freie_rg_liste erstellen -> MCS-Berechnung beschleunigen!?
        #rg_duplfrei_liste = check_rg_duplicate(mol_single_list)


        output_data = open("doku_possible_root_nodes.txt", "a")
        print "Anzahl aller duplfreien rgs", len(rg_duplfrei_liste)
        print "unique MCSs", len(all_unique_mcs_liste)

        print >> output_data, rg_data_name+".sdf"
```

```
        print >> output_data, "MCS="+str(min_atom_size)
        print >> output_data, "blacklist="+str(zinc)
        print >> output_data, "abbruch="+str(abbruch_factor*100)+"%"

        print >> output_data, "unique MCSs", len(all_unique_mcs_liste)
        print >> output_data, "Anzahl aller duplfreien rgs", len(rg_duplfrei_liste)
        output_data.close()

        print len(all_unique_mcs_liste)
        #all_unique_mcs_liste            =            determine_unique_training_mcs(mcs_matrix,
dict_duplfrei_title_2_mcs_pos,         dict_rg_cansmi_duplfrei_title,         mol_single_list,
rg_duplfrei_liste)
        #print len(all_unique_mcs_liste)


        #finde alle zur Auswahl stehenden root-nodes
        possible_root_mcs_list,                          mcs_nodes_list                          =
find_mcs_without_vorgaenger(all_unique_mcs_liste)

        output_data = open("doku_possible_root_nodes.txt", "a")
        print "Anzahl moeglicher root nodes", len(possible_root_mcs_list)
        print >> output_data, "Anzahl moeglicher root nodes", len(possible_root_mcs_list)
        output_data.close()

        #bestimme, wieviel mols zu den einzelnen root nodes gehoeren
        search_mols_belonging_to_possible_root_mcss(rg_duplfrei_liste,  possible_root_mcs_list,
mol_single_list)


        #find first root node
        used_mols_list, final_root_nodes_list = find_first_root_node(possible_root_mcs_list)
        loop_counter = 1 #gibt die Anzahl an root nodes an
        stopp_check = 0
        while stopp != 1:
            if stopp_check == 1:
                break
            #if loop_counter == 30: #hier kann man die Max-Anzahl an root-nodes steuern!
                #break
            output_data = open("doku_possible_root_nodes.txt", "a")
            print "Anzahl Gesamtmols", len(mol_single_list)
            print "Anzahl benutzte mols", len(used_mols_list)
            print >> output_data, "Anzahl Gesamtmols", len(mol_single_list)
            print >> output_data, "Anzahl benutzte mols", len(used_mols_list)
            output_data.close()

            stopp  =  determine_if_stopp_kriterium_fullfilled(mol_single_list,  used_mols_list,
abbruch_factor)
            if stopp == 1:
                print "stopp"
                break
            else:
                print "do not stopp"
                #find rest root nodes -> hier fehlt noch der Fall, dass alle root-nodes nicht
zu abbruch fuehren - geht das??? (-> erst mal nachdenken...)
                used_mols_list,          final_root_nodes_list,          stopp_check          =
find_additional_root_nodes(possible_root_mcs_list, used_mols_list, final_root_nodes_list)
            if stopp_check == 0:
                loop_counter += 1
            #all root nodes found
            #print "\n"
            #for mcsn in possible_root_mcs_list:
                #print mcsn.smile, len(mcsn.unused_ms), len(mcsn.matching_ms)

        root_counter = 0
        output_data = open("doku_possible_root_nodes.txt", "a")
        print "\nfinal root nodes, total number", loop_counter
        print >> output_data, "\nfinal root nodes, total number", loop_counter
        for final_root_node in final_root_nodes_list:
            print final_root_node.smile
            print >> output_data, final_root_node.smile
            MCS_Node.rename_node(final_root_node, "root_"+str(root_counter))
            root_counter += 1
        output_data.close()

        ##output: unused mols
        unused_mols = []
        for ms in mol_single_list:
            if ms not in used_mols_list:
```

```
                mol = ms.mol
                OESetSDData(mol, "RG_SMILE", str(ms.rg_cansmi))
                OESetSDData(mol, "P_ACTIVITY", str(ms.p_activity))
                unused_mols.append(mol)
        st.write_mols_to_sdf_gz("unused_mols", unused_mols)

        ##del vorgaenger nachfolger beziehungen und mcs-nodes mols zuordnen
        for mcsn in mcs_nodes_list:
            MCS_Node.remove_vorgaenger(mcsn)
            MCS_Node.remove_nachfolger(mcsn)
            MCS_Node.finde_mols(mcsn, rg_duplfrei_liste, mol_single_list)

        #finde alle Superstruktur-MCS (-> geht das ueber nachfolger???-> neee irgendwie
nicht...)-:
        for root in final_root_nodes_list:
        #root = final_root_nodes_list[0]
        #print "\n", root.smile, root.name

            #alle mcs, die superstruktur von root
            all_mcs_2 = []
            for mcsn_smile in all_unique_mcs_liste:
                mcsmol = OEGraphMol()
                OEParseSmiles(mcsmol, mcsn_smile)
                if root.mcsmol.NumAtoms() != mcsmol.NumAtoms() or root.smile == mcsn_smile:
#neu: 13-03-13 => wenn zum root-node noch annelierter mcs node existiert
                    match = mache_substruktur_suche(root.smile, mcsn_smile)
                    if match == 1:
                        for mcsn in mcs_nodes_list:
                            if mcsn.smile == mcsn_smile:
                                all_mcs_2.append(mcsn)

            #print len(all_mcs_2)

            network_islands_list.append(all_mcs_2)

            #beziehungen herstellen -> 26-9-12 23-30 -> juhuu, es fkt...!!! (-:
            find_mcs_relationships(all_mcs_2)


        ##Netzwerk aufbauen
        G=nx.Graph()

        G = build_mcs_network(network_islands_list, final_root_nodes_list, G)


        ##MST

        MST = find_minimum_spanning_tree(G)

        output_data = open("doku_possible_root_nodes.txt", "a")
        print >> output_data, "connected components", nx.number_connected_components(MST)
        print "connected components", nx.number_connected_components(MST)
        print >> output_data, "gesamt_mcs_nodes", MST.number_of_nodes()
        print "gesamt_mcs_nodes", MST.number_of_nodes()

        aktualisiere_vorgaenger_nachfolger_beziehung(MST)

        ##alle Molekuele an MST-MCS-Knoten visualisieren

        MST_all_mols = add_all_mols_2_mcs_network(MST)


        ##Steckenbleiben der Molekuele bestimmen
        determine_steckenbleiben_nodes(mol_single_list)

        MST_steckenbleib_mols       =       delete_edges_for_steckenbleiben_mols(MST_all_mols,
mol_single_list)


        #Netzwerkexport und Attributexport

        export_network_2_cytoscape_MST(MST_steckenbleib_mols,
"Netzwerk_MST_steckenbleib_mols_"+rg_data_name+"_"+str(date))

        export_attributes_2_cytoscape(MST_steckenbleib_mols, 0, str(date), rg_data_name)
```

Literaturverzeichnis

[1] Munos, B. Lessons from 60 Years of Pharmaceutical Innovation. *Nat. Rev. Drug Discovery* **2009,** *8* (12), 959–968.

[2] Aronson, J. K. Rare Diseases and Orphan Drugs. *Br. J. Clin. Pharmacol.* **2006,** *61* (3), 243–245.

[3] Trouiller, P.; Olliaro, P.; Torreele, E.; Orbinski, J.; Laing, R.; Ford, N. Drug Development for Neglected Diseases: A Deficient Market and a Public-Health Policy Failure. *Lancet* **2002,** *359* (9324), 2188–2194.

[4] Pécoul, B.; Chirac, P.; Trouiller, P.; Pinel, J. Access to Essential Drugs in Poor Countries: A Lost Battle? *JAMA, J. Am. Med. Assoc.* **1999,** *281* (4), 361–367.

[5] Wästfelt, M.; Fadell, B.; Henter, J.-I. A Journey of Hope: Lessons Learned from Studies on Rare Diseases and Orphan Drugs. *J. Intern. Med.* **2006,** *260* (1), 1–10.

[6] Nwaka, S.; Hudson, A. Innovative Lead Discovery Strategies for Tropical Diseases. *Nat. Rev. Drug Discovery* **2006,** *5* (11), 941–955.

[7] Brookmeyer, R.; Johnson, E.; Ziegler-Graham, K.; Arrighi, H. Michael. Forecasting the Global Burden of Alzheimer's Disease. *Alzheimers Dement.* **2007,** *3* (3), 186–191.

[8] Bray, F.; Jemal, A.; Grey, N.; Ferlay, J.; Forman, D. Global Cancer Transitions According to the Human Development Index (2008–2030): A Population-Based Study. *Lancet Oncol.* **2012,** *13* (8), 790–801.

[9] Sams-Dodd, F. Target-Based Drug Discovery: Is Something Wrong? *Drug Discovery Today* **2005,** *10* (2), 139–147.

[10] Imming, P.; Sinning, C.; Meyer, A. Drugs, their Targets and the Nature and Number of Drug Targets. *Nat. Rev. Drug Discovery* **2006,** *5* (10), 821–834.

[11] Hyman, S. E.; Fenton, W. S. What are the Right Targets for Psychopharmacology? *Science* **2003,** *299* (5605), 358–359.

[12] Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. How Many Drug Targets are There? *Nat. Rev. Drug Discovery* **2006,** *5* (12), 993–996.

[13] Hopkins, A. L.; Groom, C. R. The Druggable Genome. *Nat. Rev. Drug Discovery* **2002,** *1* (9), 727–730.

[14] Coates, A.; Hu, Y.; Bax, R.; Page, C. The Future Challenges Facing the Development of New Antimicrobial Drugs. *Nat. Rev. Drug Discovery* **2002,** *1* (11), 895–910.

[15] Wongsrichanalai, C.; Varma, J. K.; Juliano, J. J.; Kimerling, M. E.; MacArthur, J. R. Extensive Drug Resistance in Malaria and Tuberculosis. *Emerging Infect. Dis.* **2010,** *16* (7), 1063-1067.

[16] Flexner, C. HIV Drug Development: The Next 25 Years. *Nat. Rev. Drug Discovery* **2007,** *6* (12), 959–966.

[17] Szakacs, G.; Paterson, J. K.; Ludwig, J. A.; Booth-Genthe, C.; Gottesman, M. M. Targeting Multidrug Resistance in Cancer. *Nat. Rev. Drug Discovery* **2006,** *5* (3), 219–234.

[18] Kubinyi, H. Chance Favors the Prepared Mind - From Serendipity to Rational Drug Design. *J. Recept. Signal Transduction* **1999,** *19* (1-4), 15–39.

[19] Ban, T. A. The Role of Serendipity in Drug Discovery. *Dialogues Clin. Neurosci.* **2006,** *8* (3), 335–344.

[20] Hargrave-Thomas, E.; Yu, B.; Reynisson, J. Serendipity in Anticancer Drug Discovery. *World J. Clin. Oncol.* **2012,** *3* (1), 1–6.

[21] Fleming, A. On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to their Use in the Isolation of B. Influenzae. *Br. J. Exp. Pathol.* **1929,** *10* (3), 226–236.

[22] Rosenberg, B.; van Camp, L.; Krigas, T. Inhibition of Cell Division in Escherichia Coli by Electrolysis Products from a Platinum Electrode. *Nature* **1965,** *205* (4972), 698–699.

[23] Galanski, M.; Keppler, B. K. Tumorhemmende Metallverbindungen: Entwicklung, Bedeutung und Perspektiven. *Pharm. Unserer Zeit* **2006,** *35* (2), 118–123.

[24] Osterloh, I. H. The Discovery and Development of Viagra® (Sildenafil Citrate). In *Sildenafil*; Dunzendorfer, U., Ed.; Birkhäuser: Basel, 2004; Milestones in Drug Therapy, Bd. IX; S. 1-13.

[25] Michael Braun. Viagra - die Wirkung bleibt, der Umsatz schlafft ab. *http://www.heute.de/Viagra-die-Wirkung-bleibt-der-Umsatz-schlafft-ab-28494150.html.*

[26] Smith, R. A.; Griebenow, N. Combinatorial Chemistry and High-Throughput Screening. In *High-Throughput Screening in Drug Discovery*; Hüser, J., Ed.; Wiley-VCH: Weinheim, 2006; Methods and Principles in Medicinal Chemistry, Bd. 35; S. 259–296.

[27] Dobson, C. M. Chemical Space and Biology. *Nature* **2004,** *432* (7019), 824–828.

[28] Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual Screening - An Overview. *Drug Discovery Today* **1998,** *3* (4), 160–178.

[29] Lahana, R. How Many Leads from HTS? *Drug Discovery Today* **1999,** *4* (10), 447–448.

[30] Lipinski, C.; Hopkins, A. Navigating Chemical Space for Biology and Medicine. *Nature* **2004,** *432* (7019), 855–861.

[31] Aherne, G.; McDonald, E.; Workman, P. Finding the Needle in the Haystack: Why High-Throughput Screening is Good for your Health. *Breast Cancer Res.* **2002,** *4* (4), 1-7.

[32] Böhm, H.-J., Schneider, G., Eds. *Virtual Screening for Bioactive Molecules;* Wiley-VCH: Weinheim, 2000; Methods and Principles in Medicinal Chemistry, Bd. 10.

[33] Sotriffer, C., Ed. *Virtual Screening*. *Principles, Challenges, and Practical Guidelines;* Wiley-VCH: Weinheim, 2011; Methods and Principles in Medicinal Chemistry, Bd. 48.

[34] Klebe, G. *Wirkstoffdesign*. *Entwurf und Wirkung von Arzneistoffen*, 2. Auflage; Spektrum Akademischer Verlag: Heidelberg, 2009.

[35] Wermuth, C. G.; Ganellin, C. R.; Lindberg, P.; Mitscher, L. A. Glossary of Terms Used in Medicinal Chemistry (IUPAC Recommendations 1998). *Pure Appl. Chem.* **1998,** *70* (5), 1129–1143.

[36] Duffy, B. C.; Zhu, L.; Decornez, H.; Kitchen, D. B. Early Phase Drug Discovery: Cheminformatics and Computational Techniques in Identifying Lead Series. *Bioorg. Med. Chem.* **2012,** *20* (18), 5324–5342.

[37] Schnecke, V.; Boström, J. Computational Chemistry-Driven Decision Making in Lead Generation. *Drug Discovery Today* **2006,** *11* (1–2), 43–50.

[38] Bleicher, K. H.; Bohm, H.-J.; Muller, K.; Alanine, A. I. Hit and Lead Generation: Beyond High-Throughput Screening. *Nat. Rev. Drug Discovery* **2003,** *2* (5), 369–378.

[39] Schneider, G.; Fechner, U. Computer-Based De Novo Design of Drug-Like Molecules. *Nat. Rev. Drug Discovery* **2005,** *4* (8), 649–663.

[40] Nicholls, A.; Grant, J. A. Molecular Shape and Electrostatics in the Encoding of Relevant Chemical Information. *J. Comput.-Aided Mol. Des.* **2005,** *19* (9-10), 661–686.

[41] Nicholls, A.; McGaughey, G. B.; Sheridan, R. P.; Good, A. C.; Warren, G.; Mathieu, M.; Muchmore, S. W.; Brown, S. P.; Grant, J. Andrew; Haigh, J. A.; Nevins, N.; Jain, A. N.; Kelley, B. Molecular Shape and Medicinal Chemistry: A Perspective. *J. Med. Chem.* **2010,** *53* (10), 3862–3886.

[42] Bissantz, C.; Kuhn, B.; Stahl, M. A Medicinal Chemist's Guide to Molecular Interactions. *J. Med. Chem.* **2010,** *53* (14), 5061–5084.

[43] Michel, J.; Tirado-Rives, J.; Jorgensen, W. L. Energetics of Displacing Water Molecules from Protein Binding Sites: Consequences for Ligand Optimization. *J. Am. Chem. Soc.* **2009,** *131* (42), 15403–15411.

[44] Schneider, G.; Baringhaus, K.-H. *Molecular Design*. *Concepts and Applications;* Wiley-VCH: Weinheim, 2008.

[45] Ehrlich, P. Über den jetzigen Stand der Chemotherapie. *Ber. Dtsch. Chem. Ges.* **1909,** *42* (1), 17–47.

[46] Mayr, L. M.; Bojanic, D. Novel Trends in High-Throughput Screening. *Curr. Opin. Pharmacol.* **2009,** *9* (5), 580–588.

[47] Mayr, L. M.; Fuerst, P. The Future of High-Throughput Screening. *J. Biomol. Screening* **2008,** *13* (6), 443–448.

[48] Rohrer, S. G.; Baumann, K. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. *J. Chem. Inf. Model.* **2009,** *49* (2), 169–184.

[49] McGovern, S. L.; Helfand, B. T.; Feng, B.; Shoichet, B. K. A Specific Mechanism of Nonspecific Inhibition. *J. Med. Chem.* **2003,** *46* (20), 4265–4272.

[50] Seidler, J.; McGovern, S. L.; Doman, T. N.; Shoichet, B. K. Identification and Prediction of Promiscuous Aggregating Inhibitors among Known Drugs. *J. Med. Chem.* **2003,** *46* (21), 4477–4486.

[51] Feng, B. Y.; Simeonov, A.; Jadhav, A.; Babaoglu, K.; Inglese, J.; Shoichet, B. K.; Austin, C. P. A High-Throughput Screen for Aggregation-Based Inhibition in a Large Compound Library. *J. Med. Chem.* **2007,** *50* (10), 2385–2390.

[52] McGovern, S. L.; Shoichet, B. K. Kinase Inhibitors: Not Just for Kinases Anymore. *J. Med. Chem.* **2003,** *46* (8), 1478–1483.

[53] Coan, K. E. D.; Maltby, D. A.; Burlingame, A. L.; Shoichet, B. K. Promiscuous Aggregate-Based Inhibitors Promote Enzyme Unfolding. *J. Med. Chem.* **2009,** *52* (7), 2067–2075.

[54] Shoichet, B. K. Screening in a Spirit Haunted World. *Drug Discovery Today* **2006,** *11* (13–14), 607–615.

[55] Rishton, G. M. Reactive Compounds and In Vitro False Positives in HTS. *Drug Discovery Today* **1997,** *2* (9), 382–384.

[56] Ludewig, S.; Kossner, M.; Schiller, M.; Baumann, K.; Schirmeister, T. Enzyme Kinetics and Hit Validation in Fluorimetric Protease Assays. *Curr. Top. Med. Chem.* **2010,** *10* (3), 368–382.

[57] McGovern, S. L.; Caselli, E.; Grigorieff, N.; Shoichet, B. K. A Common Mechanism Underlying Promiscuous Inhibitors from Virtual and High-Throughput Screening. *J. Med. Chem.* **2002,** *45* (8), 1712–1722.

[58] Jadhav, A.; Ferreira, R. S.; Klumpp, C.; Mott, B. T.; Austin, C. P.; Inglese, J.; Thomas, C. J.; Maloney, D. J.; Shoichet, B. K.; Simeonov, A. Quantitative Analyses of Aggregation, Autofluorescence, and Reactivity Artifacts in a Screen for Inhibitors of a Thiol Protease. *J. Med. Chem.* **2009,** *53* (1), 37–51.

[59] Thorne, N.; Auld, D. S.; Inglese, J. Apparent Activity in High-Throughput Screening: Origins of Compound-Dependent Assay Interference. *Curr. Opin. Chem. Biol.* **2010,** *14* (3), 315–324.

[60] Keserű, G. M.; Makara, G. M. Hit Discovery and Hit-to-Lead Approaches. *Drug Discovery Today* **2006,** *11* (15–16), 741–748.

[61] Hüser, J.; Kalthof, B.; Strayle, J. Functional Cell-based Assays for Targeted Lead Discovery in High-Throughput Screening. In *High-Throughput Screening in Drug Discovery*; Hüser, J., Ed.; Wiley-VCH: Weinheim, 2006; Methods and Principles in Medicinal Chemistry, Bd. 35; S. 75–91.

[62] Mallender, W. D.; Bembenek, M.; Dick, L. R.; Kuranda, M.; Li, P.; Menon, S.; Pardo, E.; Parsons, T. Biochemical Assays for HighTthroughput Screening. In *High-Throughput Screening in Drug Discovery*; Hüser, J., Ed.; Wiley-VCH: Weinheim, 2006; Methods and Principles in Medicinal Chemistry, Bd. 35; S. 93–128.

[63] Parandoosh, Z. Cell-Based Assays. *J. Biomol. Screening* **1997,** *2* (4), 201–202.

[64] Moore, K.; Rees, S. Cell-Based Versus Isolated Target Screening: How Lucky Do You Feel? *J. Biomol. Screening* **2001,** *6* (2), 69–74.

[65] Manly, S. P. In Vitro Biochemical Screening. *J. Biomol. Screening* **1997,** *2* (4), 197–199.

[66] Zhu, Z.; Kim, S.; Chen, T.; Lin, J.-H.; Bell, A.; Bryson, J.; Dubaquie, Y.; Yan, N.; Yanchunas, J.; Xie, D.; Stoffel, R.; Sinz, M.; Dickinson, K. Correlation of High-Throughput Pregnane X Receptor (PXR) Transactivation and Binding Assays. *J. Biomol. Screening* **2004,** *9* (6), 533–540.

[67] Neubig, R. R.; Spedding, M.; Kenakin, T.; Christopoulos, A. International Union of Pharmacology Committee on Receptor Nomenclature and Drug Classification. XXXVIII. Update on Terms and Symbols in Quantitative Pharmacology. *Pharmacol. Rev.* **2003,** *55* (4), 597–606.

[68] Höfliger, M. M.; Beck-Sickinger, A. G. Receptor-Ligand Interaction. *Protein-Ligand Interactions*; Wiley-VCH Verlag GmbH & Co. KGaA, 2005; S. 107–135.

[69] Yung-Chi, C.; Prusoff, W. H. Relationship Between the Inhibition Constant (K_i) and the Concentration of Inhibitor Which Causes 50 Per Cent Inhibition (IC_{50}) of an Enzymatic Reaction. *Biochem. Pharmacol.* **1973,** *22* (23), 3099–3108.

[70] Bender, A. Databases: Compound Bioactivities Go Public. *Nat. Chem. Biol.* **2010,** *6* (5), 309.

[71] Wang, Y.; Bolton, E.; Dracheva, S.; Karapetyan, K.; Shoemaker, B. A.; Suzek, T. O.; Wang, J.; Xiao, J.; Zhang, J.; Bryant, S. H. An Overview of the PubChem BioAssay Resource. *Nucleic Acids Res.* **2010,** *38* (suppl 1), D255-D266.

[72] Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Zhou, Z.; Han, L.; Karapetyan, K.; Dracheva, S.; Shoemaker, B. A.; Bolton, E.; Gindulyte, A.; Bryant, S. H. PubChem's BioAssay Database. *Nucleic Acids Res.* **2012,** *40* (D1), D400-D412.

[73] Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012,** *40* (D1), D1100-D1107.

[74] ChEMBL. *http://www.ebi.ac.uk/chembldb/.*

[75] Chen, X.; Lin, Y.; Liu, M.; Gilson, M. K. The Binding Database: Data Management and Interface Design. *Bioinformatics* **2002,** *18* (1), 130–139.

[76] Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: A Web-Accessible Database of Experimentally Determined Protein–Ligand Binding Affinities. *Nucleic Acids Res.* **2007,** *35* (suppl 1), D198-D201.

[77] BindingDB. *http://www.bindingdb.org/.*

[78] Nicola, G.; Liu, T.; Gilson, M. K. Public Domain Databases for Medicinal Chemistry. *J. Med. Chem.* **2012,** *55* (16), 6987–7002.

[79] Oprea, T. I.; Tropsha, A. Target, Chemical and Bioactivity Databases – Integration Is Key. *Drug Discovery Today: Technol.* **2006,** *3* (4), 357–365.

[80] NCBI PubChem Sources. *http://pubchem.ncbi.nlm.nih.gov/sources/sources.cgi#assa/.*

[81] The EMBL-European Bioinformatics Institute. ChEMBL Database Version 17 Release Notes. *ftp://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/releases/chembl_17/chembl_17_release_notes.txt.*

[82] Wassermann, A. M.; Bajorath, J. BindingDB and ChEMBL: Online Compound Databases for Drug Discovery. *Expert Opin. Drug Discovery* **2011,** *6* (7), 683–687.

[83] Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein−Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* **2004,** *47* (12), 2977–2980.

[84] Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. The PDBbind Database: Methodologies and Updates. *J. Med. Chem.* **2005,** *48* (12), 4111–4119.

[85] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000,** *28* (1), 235–242.

[86] RCSB Protein Data Bank - RCSB PDB. *http://www.rcsb.org/.*

[87] Olah, M.; Mracec, M.; Ostopovici, L.; Rad, R.; Bora, A.; Hadaruga, N.; Olah, I.; Banda, M.; Simon, Z.; Mracec, M.; Oprea, T. I. WOMBAT: World of Molecular Bioactivity. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, 2005; Methods and Principles in Medicinal Chemistry, Bd. 23; S. 221–239.

[88] Olah, M.; Rad, R.; Ostopovici, L.; Bora, A.; Hadaruga, N.; Hadaruga, D.; Moldovan, R.; Fulias, A.; Mractc, M.; Oprea, T. I. WOMBAT and WOMBAT-PK: Bioactivity Databases for Lead and Drug Discovery. In *Chemical Biology*: *From Small Molecules to Systems Biology and Drug Design*, Bd. 2; Schreiber, S. L., Kapoor, T. M., Wess, G., Eds.; Wiley-VCH: Weinheim, 2007; S. 760–786.

[89] WOMBAT 2013.1. Sunset Molecular Discovery. *http://www.sunsetmolecular.com/index.php?option=com_content&view=article&id=15&Itemid=10.*

[90] Southan, C.; Varkonyi, P.; Muresan, S. Quantitative Assessment of the Expanding Complementarity between Public and Commercial Databases of Bioactive Compounds. *J. Cheminf.* **2009,** *1* (1), 10.

[91] Tiikkainen, P.; Franke, L. Analysis of Commercial and Public Bioactivity Databases. *J. Chem. Inf. Model.* **2012,** *52* (2), 319–326.

[92] Williams, A. J.; Ekins, S. A Quality Alert and Call for Improved Curation of Public Chemistry Databases. *Drug Discovery Today* **2011,** *16* (17–18), 747–750.

[93] Williams, A. J.; Ekins, S.; Tkachenko, V. Towards a Gold Standard: Regarding Quality in Public Domain Chemistry Databases and Approaches to Improving the Situation. *Drug Discovery Today* **2012,** *17* (13–14), 685–701.

[94] Kramer, C.; Lewis, R. QSARs, Data and Error in the Modern Age of Drug Discovery. *Curr. Top. Med. Chem.* **2012,** *12* (17), 1896–1902.

[95] Tiikkainen, P.; Bellis, L.; Light, Y.; Franke, L. Estimating Error Rates in Bioactivity Databases. *J. Chem. Inf. Model.* **2013,** *53* (10), 2499–2505.

[96] Kalliokoski, T.; Kramer, C.; Vulpetti, A. Quality Issues with Public Domain Chemogenomics Data. *Mol. Inf.* **2013,** *32* (11-12), 898–905.

[97] Young, D.; Martin, T.; Venkatapathy, R.; Harten, P. Are the Chemical Structures in Your QSAR Correct? *QSAR Comb. Sci.* **2008,** *27* (11-12), 1337–1345.

[98] Kramer, C.; Kalliokoski, T.; Gedeck, P.; Vulpetti, A. The Experimental Uncertainty of Heterogeneous Public Ki Data. *J. Med. Chem.* **2012,** *55* (11), 5165–5173.

[99] Kalliokoski, T.; Kramer, C.; Vulpetti, A.; Gedeck, P. Comparability of Mixed IC_{50} Data – A Statistical Analysis. *PLoS ONE* **2013,** *8* (4), e61007.

[100] Stumpfe, D.; Bajorath, J. Assessing the Confidence Level of Public Domain Compound Activity Data and the Impact of Alternative Potency Measurements on SAR Analysis. *J. Chem. Inf. Model.* **2011,** *51* (12), 3131–3137.

[101] Fourches, D.; Muratov, E.; Tropsha, A. Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* **2010,** *50* (7), 1189–1204.

[102] Hu, Y.; Bajorath, J. Growth of Ligand–Target Interaction Data in ChEMBL Is Associated with Increasing and Activity Measurement-Dependent Compound Promiscuity. *J. Chem. Inf. Model.* **2012,** *52* (10), 2550–2558.

[103] Johnson M. A., Maggiora G. M., Eds. *Concepts and Applications of Molecular Similarity;* John Wiley & Sons: New York, 1990.

[104] Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998,** *38* (6), 983–996.

[105] Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **2002,** *45* (19), 4350–4358.

[106] Kubinyi, H. Similarity and Dissimilarity: A Medicinal Chemist's View. *Perspect. Drug Discovery Des.* **1998,** *9-11* (0), 225–252.

[107] Leung, C. S.; Leung, S. S. F.; Tirado-Rives, J.; Jorgensen, W. L. Methyl Effects on Protein–Ligand Binding. *J. Med. Chem.* **2012,** *55* (9), 4489–4500.

[108] Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity Landscape Representations for Structure−Activity Relationship Analysis. *J. Med. Chem.* **2010,** *53* (23), 8209–8223.

[109] Todeschini, R., Consonni, V., Eds. *Molecular Descriptors for Chemoinformatics*, 2. Aufl.; Wiley-VCH: Weinheim, 2009; Methods and Principles in Medicinal Chemistry, Bd. 41.

[110] Bajorath, J. Molecular Crime Scene Investigation – Dusting for Fingerprints. *Drug Discovery Today: Technol.* **2013,** *10* (4), e491–e498.

[111] MACCS Structural Keys. Symyx Technologies, Inc., Sunnyvale, CA. *http://www.symyx.com/.*

[112] Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002,** *42* (6), 1273–1280.

[113] James, C. A.; Weininger, D. *Daylight Theory Manual;* Daylight Chemical Information Systems: Los Altos, 2006.

[114] O'Boyle, N.; Banck, M.; James, C.; Morley, C.; Vandermeersch, T.; Hutchison, G. Open Babel: An Open Chemical Toolbox. *J. Cheminf.* **2011,** *3* (1), 33.

[115] Open Babel 2.3.1; *http://openbabel.org/.*

[116] Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010,** *50* (5), 742–754.

[117] Pickett, S. D.; Mason, J. S.; McLay, I. M. Diversity Profiling and Design Using 3D Pharmacophores: Pharmacophore-Derived Queries (PDQ). *J. Chem. Inf. Comput. Sci.* **1996,** *36* (6), 1214–1223.

[118] Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-Point Pharmacophore Method for Molecular Similarity and Diversity Applications: Overview of the Method and Applications, Including a Novel Approach to the Design of Combinatorial Libraries Containing Privileged Substructures. *J. Med. Chem.* **1999,** *42* (17), 3251–3264.

[119] Deng, Z.; Chuaqui, C.; Singh, J. Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein−Ligand Binding Interactions. *J. Med. Chem.* **2003,** *47* (2), 337–344.

[120] Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*, überarb. Aufl.; Springer: Dordrecht, 2007.

[121] Scior, T.; Bender, A.; Tresadern, G.; Medina-Franco, J. L.; Martínez-Mayorga, K.; Langer, T.; Cuanalo-Contreras, K.; Agrafiotis, D. K. Recognizing Pitfalls in Virtual Screening: A Critical Review. *J. Chem. Inf. Model.* **2012,** *52* (4), 867–881.

[122] Sheridan, R. P.; Kearsley, S. K. Why Do We Need So Many Chemical Similarity Search Methods? *Drug Discovery Today* **2002,** *7* (17), 903–911.

[123] Weininger, D. SMILES, A Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988,** *28* (1), 31–36.

[124] Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989,** *29* (2), 97–101.

[125] The IUPAC International Chemical Identifier (InChI). *http://www.iupac.org/home/publications/e-resources/inchi.html.*

[126] The InChI Trust. *http://www.inchi-trust.org/.*

[127] Thalheim, T.; Vollmer, A.; Ebert, R.-U.; Kühne, R.; Schüürmann, G. Tautomer Identification and Tautomer Structure Generation Based on the InChI Code. *J. Chem. Inf. Model.* **2010,** *50* (7), 1223–1232.

[128] Downs, G. M.; Willett, P. Similarity Searching in Databases of Chemical Structures. *Reviews in Computational Chemistry*; John Wiley & Sons, Inc, 1995; S. 1–66.

[129] Cao, Y.; Jiang, T.; Girke, T. A Maximum Common Substructure-based Algorithm for Searching and Predicting Drug-like Compounds. *Bioinformatics* **2008,** *24* (13), i366-i374.

[130] Raymond, J. W.; Gardiner, E. J.; Willett, P. RASCAL: Calculation of Graph Similarity Using Maximum Common Edge Subgraphs. *Comput. J.* **2002,** *45* (6), 631–644.

[131] Raymond, J. W.; Gardiner, E. J.; Willett, P. Heuristics for Similarity Searching of Chemical Graphs Using a Maximum Common Edge Subgraph Algorithm. *J. Chem. Inf. Comput. Sci.* **2002,** *42* (2), 305–316.

[132] Raymond, J. W.; Willett, P. Maximum Common Subgraph Isomorphism Algorithms for the Matching of Chemical Structures. *J. Comput.-Aided Mol. Des.* **2002,** *16* (7), 521–533.

[133] Raymond, J.; Willett, P. Effectiveness of Graph-Based and Fingerprint-Based Similarity Measures for Virtual Screening of 2D Chemical Structure Databases. *J. Comput.-Aided Mol. Des.* **2002,** *16* (1), 59-71.

[134] Maggiora, G. M. On Outliers and Activity Cliffs - Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006,** *46* (4), 1535-1535.

[135] Peltason, L.; Bajorath, J. SAR Index: Quantifying the Nature of Structure−Activity Relationships. *J. Med. Chem.* **2007,** *50* (23), 5571–5578.

[136] Lounkine, E.; Wawer, M.; Wassermann, A. M.; Bajorath, J. SARANEA: A Freely Available Program To Mine Structure−Activity and Structure−Selectivity Relationship Information in Compound Data Sets. *J. Chem. Inf. Model.* **2010,** *50* (1), 68–78.

[137] Guha, R.; van Drie, J. H. Structure−Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model.* **2008,** *48* (3), 646–658.

[138] Hu, Y.; Maggiora, G.; Bajorath, J. Activity Cliffs in PubChem Confirmatory Bioassays Taking Inactive Compounds into Account. *J. Comput.-Aided Mol. Des.* **2013,** *27* (2), 115-124.

[139] Hu, Y.; Bajorath, J. Extending the Activity Cliff Concept: Structural Categorization of Activity Cliffs and Systematic Identification of Different Types of Cliffs in the ChEMBL Database. *J. Chem. Inf. Model.* **2012,** *52* (7), 1806–1811.

[140] Stumpfe, D.; Bajorath, J. Frequency of Occurrence and Potency Range Distribution of Activity Cliffs in Bioactive Compounds. *J. Chem. Inf. Model.* **2012,** *52* (9), 2348–2353.

[141] Vogt, M.; Huang, Y.; Bajorath, J. From Activity Cliffs to Activity Ridges: Informative Data Structures for SAR Analysis. *J. Chem. Inf. Model.* **2011,** *51* (8), 1848–1856.

[142] Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *J. Med. Chem.* **2012,** *55* (7), 2932–2942.

[143] Stumpfe, D.; Hu, Y.; Dimova, D.; Bajorath, J. Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry. *J. Med. Chem.* **2013,** *57* (1), 18–28.

[144] Guha, R. Exploring Uncharted Territories: Predicting Activity Cliffs in Structure–Activity Landscapes. *J. Chem. Inf. Model.* **2012,** *52* (8), 2181–2191.

[145] Heikamp, K.; Hu, X.; Yan, A.; Bajorath, J. Prediction of Activity Cliffs Using Support Vector Machines. *J. Chem. Inf. Model.* **2012,** *52* (9), 2354–2365.

[146] Namasivayam, V.; Iyer, P.; Bajorath, J. Prediction of Individual Compounds Forming Activity Cliffs Using Emerging Chemical Patterns. *J. Chem. Inf. Model.* **2013,** *53* (12), 3131–3139.

[147] Johnson, S. R. The Trouble with QSAR (or How I Learned To Stop Worrying and Embrace Fallacy). *J. Chem. Inf. Model.* **2007,** *48* (1), 25–26.

[148] Medina-Franco, J. L. Activity Cliffs: Facts or Artifacts? *Chem. Biol. Drug Des.* **2013,** *81* (5), 553–556.

[149] Yongye, A. B.; Byler, K.; Santos, R.; Martínez-Mayorga, K.; Maggiora, G. M.; Medina-Franco, J. L. Consensus Models of Activity Landscapes with Multiple Chemical, Conformer, and Property Representations. *J. Chem. Inf. Model.* **2011,** *51* (6), 1259–1270.

[150] Medina-Franco, J. L.; Martínez-Mayorga, K.; Bender, A.; Marín, R. M.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A. Characterization of Activity Landscapes Using 2D and 3D Similarity Methods: Consensus Activity Cliffs. *J. Chem. Inf. Model.* **2009,** *49* (2), 477–491.

[151] Hu, X.; Hu, Y.; Vogt, M.; Stumpfe, D.; Bajorath, J. MMP-Cliffs: Systematic Identification of Activity Cliffs on the Basis of Matched Molecular Pairs. *J. Chem. Inf. Model.* **2012,** *52* (5), 1138–1145.

[152] Seebeck, B.; Wagener, M.; Rarey, M. From Activity Cliffs to Target-Specific Scoring Models and Pharmacophore Hypotheses. *ChemMedChem* **2011,** *6* (9), 1630-1639.

[153] Langdon, S. R.; Ertl, P.; Brown, N. Bioisosteric Replacement and Scaffold Hopping in Lead Generation and Optimization. *Mol. Inf.* **2010,** *29* (5), 366–385.

[154] Meanwell, N. A. Synopsis of Some Recent Tactical Application of Bioisosteres in Drug Design. *J. Med. Chem.* **2011,** *54* (8), 2529–2591.

[155] Brown, N., Ed. *Bioisosteres in Medicinal Chemistry;* Wiley-VCH: Weinheim, 2012; Methods and Principles in Medicinal Chemistry, Bd. 54.

[156] Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J. Exploration of Structure−Activity Relationship Determinants in Analogue Series. *J. Med. Chem.* **2009,** *52* (10), 3212–3224.

[157] Wassermann, A. M.; Peltason, L.; Bajorath, J. Computational Analysis of Multi-Target Structure–Activity Relationships to Derive Preference Orders for Chemical Modifications toward Target Selectivity. *ChemMedChem* **2010,** *5* (6), 847-858.

[158] Wawer, M.; Bajorath, J. Local Structural Changes, Global Data Views: Graphical Substructure−Activity Relationship Trailing. *J. Med. Chem.* **2011,** *54* (8), 2944–2951.

[159] Stumpfe, D.; Bajorath, J. Methods for SAR Visualization. *RSC Adv.* **2012,** *2* (2), 369–378.

[160] Bajorath, J.; Peltason, L.; Wawer, M.; Guha, R.; Lajiness, M. S.; van Drie, J. H. Navigating Structure–Activity Landscapes. *Drug Discovery Today* **2009,** *14* (13–14), 698–705.

[161] Wawer, M.; Lounkine, E.; Wassermannm Anne M.; Bajorath, J. Data Structures and Computational Tools for the Extraction of SAR Information from Large Compound Sets. *Drug Discovery Today* **2010,** *15* (15–16), 630–639.

[162] Bajorath, J. Large-Scale SAR Analysis. *Drug Discovery Today: Technol.* **2013,** *10* (3), e419.

[163] Hansch, C.; Fujita, T. p-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964,** *86* (8), 1616–1626.

[164] Free, S. M.; Wilson, J. W. A Mathematical Contribution to Structure-Activity Studies. *J. Med. Chem.* **1964,** *7* (4), 395–399.

[165] Kubinyi, H. From Narcosis to Hyperspace: The History of QSAR. *Quant. Struct.-Act. Relat.* **2002,** *21* (4), 348–356.

[166] Kubinyi, H. QSAR and 3D QSAR in Drug Design Part 1: Methodology. *Drug Discovery Today* **1997,** *2* (11), 457–467.

[167] Kubinyi, H. QSAR and 3D QSAR in Drug Design Part 2: Applications and Problems. *Drug Discovery Today* **1997,** *2* (12), 538–546.

[168] Agrafiotis, D. K.; Shemanarev, M.; Connolly, P. J.; Farnum, M.; Lobanov, V. S. SAR Maps: A New SAR Visualization Technique for Medicinal Chemists. *J. Med. Chem.* **2007,** *50* (24), 5926–5937.

[169] Kolpak, J.; Connolly, P. J.; Lobanov, V. S.; Agrafiotis, D. K. Enhanced SAR Maps: Expanding the Data Rendering Capabilities of a Popular Medicinal Chemistry Tool. *J. Chem. Inf. Model.* **2009,** *49* (10), 2221–2230.

[170] Böhm, H.-J.; Flohr, A.; Stahl, M. Scaffold Hopping. *Drug Discovery Today: Technol.* **2004,** *1* (3), 217–224.

[171] Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996,** *39* (15), 2887–2893.

[172] Xu, Y.-J.; Johnson, M. Algorithm for Naming Molecular Equivalence Classes Represented by Labeled Pseudographs. *J. Chem. Inf. Comput. Sci.* **2001,** *41* (1), 181–185.

[173] Langdon, S. R.; Brown, N.; Blagg, J. Scaffold Diversity of Exemplified Medicinal Chemistry Space. *J. Chem. Inf. Model.* **2011,** *51* (9), 2174–2185.

[174] Nicolaou, C. A.; Tamura, S. Y.; Kelley, B. P.; Bassett, S. I.; Nutt, R. F. Analysis of Large Screening Data Sets via Adaptively Grown Phylogenetic-Like Trees. *J. Chem. Inf. Comput. Sci.* **2002,** *42* (5), 1069–1079.

[175] Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. „Grundgerüstwechsel" (Scaffold-Hopping) durch topologische Pharmakophorsuche: ein Beitrag zum virtuellen Screening. *Angew. Chem., Int. Ed.* **1999,** *111* (19), 3068–3070.

[176] Brown, N.; Jacoby, E. On Scaffolds and Hopping in Medicinal Chemistry. *Mini-Rev. Med. Chem.* **2006,** *6* (11), 1217–1229.

[177] Brown, N., Ed. *Scaffold Hopping in Medicinal Chemistry;* Wiley-VCH: Weinheim, 2013; Methods and Principles in Medicinal Chemistry, Bd. 58.

[178] Hu, Y.; Stumpfe, D.; Bajorath, J. Lessons Learned from Molecular Scaffold Analysis. *J. Chem. Inf. Model.* **2011,** *51* (8), 1742–1753.

[179] Schuffenhauer, A.; Varin, T. Rule-Based Classification of Chemical Structures by Scaffold. *Mol. Inf.* **2011,** *30* (8), 646–664.

[180] Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The Scaffold Tree − Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. *J. Chem. Inf. Model.* **2007,** *47* (1), 47–58.

[181] Wilkens, S. J.; Janes, J.; Su, A. I. HierS: Hierarchical Scaffold Clustering Using Topological Chemical Graphs. *J. Med. Chem.* **2005,** *48* (9), 3182–3193.

[182] Cho, S.; Sun, Y. Visual Exploration of Structure-Activity Relationship Using Maximum Common Framework. *J. Comput.-Aided Mol. Des.* **2008,** *22* (8), 571–578.

[183] Varin, T.; Schuffenhauer, A.; Ertl, P.; Renner, S. Mining for Bioactive Scaffolds with Scaffold Networks: Improved Compound Set Enrichment from Primary Screening Data. *J. Chem. Inf. Model.* **2011,** *51* (7), 1528–1538.

[184] Renner, S.; van Otterlo, Willem A L; Dominguez Seoane, M.; Mocklinghoff, S.; Hofmann, B.; Wetzel, S.; Schuffenhauer, A.; Ertl, P.; Oprea, T. I.; Steinhilber, D.; Brunsveld, L.; Rauh, D.; Waldmann, H. Bioactivity-Guided Mapping and Navigation of Chemical Space. *Nat. Chem. Biol.* **2009,** *5* (8), 585–592.

[185] Wetzel, S.; Klein, K.; Renner, S.; Rauh, D.; Oprea, T. I.; Mutzel, P.; Waldmann, H. Interactive Exploration of Chemical Space with Scaffold Hunter. *Nat. Chem. Biol.* **2009,** *5* (8), 581–583.

[186] Agrafiotis, D. K.; Wiener, J. J. M. Scaffold Explorer: An Interactive Tool for Organizing and Mining Structure−Activity Data Spanning Multiple Chemotypes. *J. Med. Chem.* **2010,** *53* (13), 5002–5011.

[187] Clark, A. M.; Labute, P. Detection and Assignment of Common Scaffolds in Project Databases of Lead Molecules. *J. Med. Chem.* **2009,** *52* (2), 469–483.

[188] Molecular Operating Environment (MOE) 2012.10. Chemical Computing Group, Montreal, Canada. *http://www.chemcomp.com/.*

[189] Gupta-Ostermann, D.; Hu, Y.; Bajorath, J. Introducing the LASSO Graph for Compound Data Set Representation and Structure–Activity Relationship Analysis. *J. Med. Chem.* **2012,** *55* (11), 5546–5553.

[190] Shanmugasundaram, V.; Maggiora, G. M. Characterizing Property and Activity Landscapes Using an Information-Theoretic Approach. In *222nd ACS National Meeting*; Division of Chemical Information (American Chemical Society), Ed.: Washington DC, United States, 2001.

[191] Sukumar, N.; Krein, M. P. Graphs and Networks in Chemical and Biological Informatics: Past, Present and Future. *Future Med. Chem.* **2012,** *4* (16), 2039–2047.

[192] Hasan, S.; Bonde, B. K.; Buchan, N. S.; Hall, M. D. Network Analysis Has Diverse Roles in Drug Discovery. *Drug Discovery Today* **2012,** *17* (15–16), 869–874.

[193] Wawer, M.; Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J. Structure−Activity Relationship Anatomy by Network-like Similarity Graphs and Local Structure−Activity Relationship Indices. *J. Med. Chem.* **2008,** *51* (19), 6075–6084.

[194] Wawer, M.; Peltason, L.; Bajorath, J. Elucidation of Structure−Activity Relationship Pathways in Biological Screening Data. *J. Med. Chem.* **2009,** *52* (4), 1075–1080.

[195] Wawer, M.; Bajorath, J. Extracting SAR Information from a Large Collection of Anti-Malarial Screening Hits by NSG-SPT Analysis. *ACS Med. Chem. Lett.* **2011,** *2* (3), 201–206.

[196] Dijkstra, E. W. A Note on Two Problems in Connexion with Graphs. *Numer. Math.* **1959,** *1* (1), 269-271.

[197] Wawer, M.; Bajorath, J. Systematic Extraction of Structure–Activity Relationship Information from Biological Screening Data. *ChemMedChem* **2009,** *4* (9), 1431-1438.

[198] Wawer, M.; Sun, S.; Bajorath, J. Computational Characterization of SAR Microenvironments in High-Throughput Screening Data. *Int. J. High Throughput Screening* **2010,** *1*, 15–27.

[199] Wawer, M.; Bajorath, J. Similarity−Potency Trees: A Method to Search for SAR Information in Compound Data Sets and Derive SAR Rules. *J. Chem. Inf. Model.* **2010,** *50* (8), 1395–1409.

[200] Gamo, F.-J.; Sanz, L. M.; Vidal, J.; Cozar, C. de; Alvarez, E.; Lavandera, J.-L.; Vanderwall, D. E.; Green, Darren V. S.; Kumar, V.; Hasan, S.; Brown, J. R.; Peishoff, C. E.; Cardon, L. R.; Garcia-Bustos, J. F. Thousands of Chemical Starting Points for Antimalarial Lead Identification. *Nature* **2010,** *465* (7296), 305–310.

[201] Harper, G.; Bravi, G. S.; Pickett, S. D.; Hussain, J.; Green, D. V. S. The Reduced Graph Descriptor in Virtual Screening and Data-Driven Clustering of High-Throughput Screening Data. *J. Chem. Inf. Comput. Sci.* **2004,** *44* (6), 2145–2156.

[202] Leach, A. G.; Jones, H. D.; Cosgrove, D. A.; Kenny, P. W.; Ruston, L.; MacFaul, P.; Wood, J. Matthew; Colclough, N.; Law, B. Matched Molecular Pairs as a Guide in the Optimization of Pharmaceutical Properties; a Study of Aqueous Solubility, Plasma Protein Binding and Oral Exposure. *J. Med. Chem.* **2006,** *49* (23), 6672–6682.

[203] Griffen, E.; Leach, A. G.; Robb, G. R.; Warner, D. J. Matched Molecular Pairs as a Medicinal Chemistry Tool. *J. Med. Chem.* **2011,** *54* (22), 7739–7750.

[204] Papadatos, G.; Alkarouri, M.; Gillet, V. J.; Willett, P.; Kadirkamanathan, V.; Luscombe, C. N.; Bravi, G.; Richmond, N. J.; Pickett, S. D.; Hussain, J.; Pritchard, J. M.; Cooper, Anthony W. J.; Macdonald, Simon J. F. Lead Optimization Using Matched Molecular Pairs: Inclusion of Contextual Information for Enhanced Prediction of hERG Inhibition, Solubility, and Lipophilicity. *J. Chem. Inf. Model.* **2010,** *50* (10), 1872–1886.

[205] Wassermann, A. M.; Bajorath, J. A Data Mining Method to Facilitate SAR Transfer. *J. Chem. Inf. Model.* **2011,** *51* (8), 1857–1866.

[206] Warner, D. J.; Bridgland-Taylor, M. H.; Sefton, C. E.; Wood, D. J. Prospective Prediction of Antitarget Activity by Matched Molecular Pairs Analysis. *Mol. Inf.* **2012,** *31* (5), 365–368.

[207] Wassermann, A. M.; Dimova, D.; Iyer, P.; Bajorath, J. Advances in Computational Medicinal Chemistry: Matched Molecular Pair Analysis. *Drug Dev. Res.* **2012,** *73* (8), 518–527.

[208] Sheridan, R. P. The Most Common Chemical Replacements in Drug-Like Compounds. *J. Chem. Inf. Comput. Sci.* **2002,** *42* (1), 103–108.

[209] Sheridan, R. P.; Hunt, P.; Culberson, J. C. Molecular Transformations as a Way of Finding and Exploiting Consistent Local QSAR. *J. Chem. Inf. Model.* **2005,** *46* (1), 180–192.

[210] Warner, D. J.; Griffen, E. J.; St-Gallay, S. A. WizePairZ: A Novel Algorithm to Identify, Encode, and Exploit Matched Molecular Pairs with Unspecified Cores in Medicinal Chemistry. *J. Chem. Inf. Model.* **2010,** *50* (8), 1350–1357.

[211] Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model.* **2010,** *50* (3), 339–348.

[212] Barker, E. J.; Gardiner, E. J.; Gillet, V. J.; Kitts, P.; Morris, J. Further Development of Reduced Graphs for Identifying Bioactive Compounds. *J. Chem. Inf. Comput. Sci.* **2003,** *43* (2), 346–356.

[213] Birchall, K.; Gillet, V. J.; Harper, G.; Pickett, S. D. Evolving Interpretable Structure−Activity Relationships. 1. Reduced Graph Queries. *J. Chem. Inf. Model.* **2008,** *48* (8), 1543–1557.

[214] Birchall, K.; Gillet, V. J.; Harper, G.; Pickett, S. D. Evolving Interpretable Structure−Activity Relationship Models. 2. Using Multiobjective Optimization To Derive Multiple Models. *J. Chem. Inf. Model.* **2008,** *48* (8), 1558–1570.

[215] Birchall, K.; Gillet, V. J. Reduced Graphs and Their Applications in Chemoinformatics. In *Chemoinformatics and Computational Chemical Biology*; Bajorath, J., Ed.; Humana Press: New York, 2011; Methods in Molecular Biology, Bd. 672; S. 197–212.

[216] Birchall, K.; Gillet, V. J.; Willett, P.; Ducrot, P.; Luttmann, C. Use of Reduced Graphs To Encode Bioisosterism for Similarity-Based Virtual Screening. *J. Chem. Inf. Model.* **2009,** *49* (6), 1330–1346.

[217] Ujváry, I. Extended Summary: BIOSTER - A Database of Structurally Analogous Compounds. *Pestic. Sci.* **1997,** *51* (1), 92–95.

[218] Diestel, R. *Graphentheorie*, 4. Aufl; Springer: Berlin, 2010.

[219] Tittmann, P. *Graphentheorie: Eine anwendungsorientierte Einführung*, 1. Aufl.; Fachbuchverlag Leipzig im Carl-Hanser-Verlag: München, 2003.

[220] Turau, V. *Algorithmische Graphentheorie*, 3., überarb. Aufl.; Oldenbourg: München, 2009.

[221] Bron, C.; Kerbosch, J. Algorithm 457: Finding All Cliques of an Undirected Graph. *Commun. ACM* **1973,** *16* (9), 575–577.

[222] Prim, R. C. Shortest Connection Networks and Some Generalizations. *Bell Syst. Tech. J.* **1957,** *36* (6), 1389–1401.

[223] Kruskal, J. B. On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. *Proc. Am. Math. Soc.* **1956,** *7* (1), 48–50.

[224] Shearer, K.; Bunke, H.; Venkatesh, S. Video Indexing and Similarity Retrieval by Largest Common Subgraph Detection Using Decision Trees. *Pattern Recogn.* **2001,** *34* (5), 1075–1091.

[225] Conte, D.; Foggia, P.; Sansone, C.; Vento, M. Thirty Years of Graph Matching in Pattern Recognition. *Int. J. Pattern Recogn. Artif. Intell.* **2004,** *18* (03), 265–298.

[226] Horaud, R.; Skordas, T. Stereo Correspondence Through Feature Grouping and Maximal Cliques. *IEEE Trans. Pattern Anal. Mach. Intell.* **1989,** *11* (11), 1168–1180.

[227] Pelillo, M.; Siddiqi, K.; Zucker, S. W. Matching Hierarchical Structures Using Association graphs. *IEEE Trans. Pattern Anal. Mach. Intell.* **1999,** *21* (11), 1105–1120.

[228] Chen, L.; Robien, W. Application of the Maximal Common Substructure Algorithm to Automatic Interpretation of 13C-NMR Spectra. *J. Chem. Inf. Comput. Sci.* **1994,** *34* (4), 934–941.

[229] Stahl, M.; Mauser, H. Database Clustering with a Combination of Fingerprint and Maximum Common Substructure Methods. *J. Chem. Inf. Model.* **2005,** *45* (3), 542–548.

[230] Böcker, A. Toward an Improved Clustering of Large Data Sets Using Maximum Common Substructures and Topological Fingerprints. *J. Chem. Inf. Model.* **2008,** *48* (11), 2097–2107.

[231] LibMCS. ChemAxon, Budapest, Ungarn. *http://www.chemaxon.com/jchem/doc/user/LibMCS.html.*

[232] Gardiner, E. J.; Gillet, V. J.; Willett, P.; Cosgrove, D. A. Representing Clusters Using a Maximum Common Edge Substructure Algorithm Applied to Reduced Graphs and Molecular Graphs. *J. Chem. Inf. Model.* **2007,** *47* (2), 354–366.

[233] Willett, P. Matching of Chemical and Biological Structures Using Subgraph and Maximal Common Subgraph Isomorphism Algorithms. In *Rational Drug Design*, 108; Truhlar, D., Howe, W., Hopfinger, A., Blaney, J., Dammkoehler, R., Eds.; Springer New York, 1999; The IMA Volumes in Mathematics and its Applications; S. 11–38.

[234] Hariharan, R.; Janakiraman, A.; Nilakantan, R.; Singh, B.; Varghese, S.; Landrum, G.; Schuffenhauer, A. MultiMCS: A Fast Algorithm for the Maximum Common Substructure Problem on Multiple Molecules. *J. Chem. Inf. Model.* **2011,** *51* (4), 788–806.

[235] Gillet, V. J.; Willett, P.; Bradshaw, J. Similarity Searching Using Reduced Graphs. *J. Chem. Inf. Comput. Sci.* **2003,** *43* (2), 338–345.

[236] Gillet, V. J.; Downs, G. M.; Holliday, J. D.; Lynch, M. F.; Dethlefsen, W. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 13. Reduced Graph Generation. *J. Chem. Inf. Comput. Sci.* **1991,** *31* (2), 260–270.

[237] Rarey, M.; Dixon, J. S. Feature Trees: A New Molecular Similarity Measure Based on Tree Matching. *J. Comput.-Aided Mol. Des.* **1998,** *12* (5), 471–490.

[238] Stiefl, N.; Zaliani, A. A Knowledge-Based Weighting Approach to Ligand-Based Virtual Screening. *J. Chem. Inf. Model.* **2006,** *46* (2), 587–596.

[239] Stiefl, N.; Watson, I. A.; Baumann, K.; Zaliani, A. ErG: 2D Pharmacophore Descriptions for Scaffold Hopping. *J. Chem. Inf. Model.* **2005,** *46* (1), 208–220.

[240] Barker, E. J.; Buttar, D.; Cosgrove, D. A.; Gardiner, E. J.; Kitts, P.; Willett, P.; Gillet, V. J. Scaffold Hopping Using Clique Detection Applied to Reduced Graphs. *J. Chem. Inf. Model.* **2006,** *46* (2), 503–511.

[241] Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical Similarity Using Physiochemical Property Descriptors. *J. Chem. Inf. Comput. Sci.* **1996,** *36* (1), 118–127.

[242] Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method Using an Incremental Construction Algorithm. *J. Mol. Biol.* **1996,** *261* (3), 470–489.

[243] Brown, N. Molecular Topology. In *Bioisosteres in Medicinal Chemistry*; Brown, N., Ed.; Wiley-VCH: Weinheim, 2012; Methods and Principles in Medicinal Chemistry, Bd. 54; S. 141–153.

[244] Renner, S.; Fechner, U.; Schneider, G. Alignment-Free Pharmacophore Patterns – A Correlation-Vector Approach. In *Pharmacophores and Pharmacophore Searches*; Langer, T., Hoffmann, R. D., Eds.; Wiley-VCH: Weinheim, 2006; Methods and Principles in Medicinal Chemistry, Bd. 32; S. 49–79.

[245] Renner, S.; Schneider, G. Scaffold-Hopping Potential of Ligand-Based Similarity Concepts. *ChemMedChem* **2006,** *1* (2), 181–185.

[246] Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity Metrics for Ligands Reflecting the Similarity of the Target Proteins. *J. Chem. Inf. Comput. Sci.* **2002,** *43* (2), 391–405.

[247] Wagener, M.; Lommerse, Jos P. M. The Quest for Bioisosteric Replacements. *J. Chem. Inf. Model.* **2006,** *46* (2), 677–685.

[248] Holliday, J. D.; Jelfs, S. P.; Willett, P.; Gedeck, P. Calculation of Intersubstituent Similarity Using R-Group Descriptors. *J. Chem. Inf. Comput. Sci.* **2003,** *43* (2), 406–411.

[249] Böhm, H.-J.; Klebe, G. What Can We Learn from Molecular Recognition in Protein–Ligand Complexes for the Design of New Drugs? *Angew. Chem., Int. Ed. Engl.* **1996,** *35* (22), 2588–2614.

[250] Gohlke, H.; Klebe, G. Ansätze zur Beschreibung und Vorhersage der Bindungsaffinität niedermolekularer Liganden an makromolekulare Rezeptoren. *Angew. Chem., Int. Ed.* **2002,** *114* (15), 2764–2798.

[251] Babine, R. E.; Bender, S. L. Molecular Recognition of Protein−Ligand Complexes: Applications to Drug Design. *Chem. Rev.* **1997,** *97* (5), 1359–1472.

[252] Böhm, H.-J.; Schneider, G. *Protein-Ligand Interactions: From Molecular Recognition to Drug Design*; Methods and Principles in Medicinal Chemistry, Bd. 19; Wiley-VCH: Weinheim, 2003.

[253] Arunan, E.; Desiraju, G. R.; Klein, R. A.; Sadlej, J.; Scheiner, S.; Alkorta, I.; Clary, D. C.; Crabtree, R. H.; Dannenberg, J. J.; Hobza, P.; Kjaergaard, H. G.; Legon, A. C.; Mennucci, B.; Nesbitt, D. J. Definition of the Hydrogen Bond (IUPAC Recommendations 2011). *Pure Appl. Chem.* **2011,** *83* (8), 1637–1641.

[254] Arunan, E.; Desiraju, G. R.; Klein, R. A.; Sadlej, J.; Scheiner, S.; Alkorta, I.; Clary, D. C.; Crabtree, R. H.; Dannenberg, J. J.; Hobza, P.; Kjaergaard, H. G.; Legon, A. C.; Mennucci, B.; Nesbitt, D. J. Defining the Hydrogen Bond: An Account (IUPAC Technical Report). *Pure Appl. Chem.* **2011,** *83* (8), 1619–1636.

[255] Kubinyi, H. Hydrogen Bonding: The Last Mystery in Drug Design? In *Pharmacokinetic Optimization in Drug Research: Biological, Physicochemical, and Computational Strategies*; Testa, B., van de Waterbeemd, H., Folkers, G., Guy, R., Eds.; Verlag Helvetica Chimica Acta: Zürich, 2007; S. 513–524.

[256] Davis, A. M.; Teague, S. J. Hydrogen Bonding, Hydrophobic Interactions, and Failure of the Rigid Receptor Hypothesis. *Angew. Chem., Int. Ed.* **1999,** *38* (6), 736–749.

[257] Bruno, I.; Cole, J.; Lommerse, J. M.; Rowland, R. S.; Taylor, R.; Verdonk, M. IsoStar: A Library of Information about Nonbonded Interactions. *J. Comput.-Aided Mol. Des.* **1997,** *11* (6), 525-537.

[258] Verdonk, M. L.; Cole, J. C.; Taylor, R. SuperStar: A Knowledge-based Approach for Identifying Interaction Sites in Proteins. *J. Mol. Biol.* **1999,** *289* (4), 1093–1108.

[259] Hendlich, M.; Bergner, A.; Günther, J.; Klebe, G. Relibase: Design and Development of a Database for Comprehensive Analysis of Protein–Ligand Interactions. *J. Mol. Biol.* **2003,** *326* (2), 607–620.

[260] Allen, F. H. The Cambridge Structural Database: A Quarter of a Million Crystal Structures and Rising. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002,** *58* (3), 380–388.

[261] Sarkhel, S.; Desiraju, G. R. N-H…O, O-H…O, and C-H…O Hydrogen Bonds in Protein–Ligand Complexes: Strong and Weak Interactions in Molecular Recognition. *Proteins: Struct., Funct., Bioinf.* **2004,** *54* (2), 247–259.

[262] Panigrahi, S. K.; Desiraju, G. R. Strong and Weak Hydrogen Bonds in the Protein–Ligand Interface. *Proteins: Struct., Funct., Bioinf.* **2007,** *67* (1), 128–141.

[263] Burley, S. K.; Petsko, G. A. Aromatic-Aromatic Interaction: A Mechanism of Protein Structure Stabilization. *Science* **1985,** *229* (4708), 23–28.

[264] Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M.; Tanabe, K. Origin of Attraction and Directionality of the π/π Interaction: Model Chemistry Calculations of Benzene Dimer Interaction. *J. Am. Chem. Soc.* **2001,** *124* (1), 104–112.

[265] Dougherty, D. A. Cation-π Interactions in Chemistry and Biology: A New View of Benzene, Phe, Tyr, and Trp. *Science* **1996,** *271* (5246), 163–168.

[266] Gallivan, J. P.; Dougherty, D. A. Cation-π Interactions in Structural Biology. *Proc. Natl. Acad. Sci. U. S. A.* **1999,** *96* (17), 9459–9464.

[267] Zacharias, N.; Dougherty, D. A. Cation–π Interactions in Ligand Recognition and Catalysis. *Trends Pharmacol. Sci.* **2002,** *23* (6), 281–287.

[268] Purser, S.; Moore, P. R.; Swallow, S.; Gouverneur, V. Fluorine in Medicinal Chemistry. *Chem. Soc. Rev.* **2008,** *37* (2), 320–330.

[269] Hagmann, W. K. The Many Roles for Fluorine in Medicinal Chemistry. *J. Med. Chem.* **2008,** *51* (15), 4359–4369.

[270] Politzer, P.; Murray, J. S.; Clark, T. Halogen Bonding: An Electrostatically-Driven Highly Directional Noncovalent Interaction. *Phys. Chem. Chem. Phys.* **2010,** *12* (28), 7748–7757.

[271] Lu, Y.; Shi, T.; Wang, Y.; Yang, H.; Yan, X.; Luo, X.; Jiang, H.; Zhu, W. Halogen Bonding-A Novel Interaction for Rational Drug Design? *J. Med. Chem.* **2009,** *52* (9), 2854–2862.

[272] Kolar, M.; Hobza, P.; Bronowska, A. K. Plugging the Explicit [Sigma]-Holes in Molecular Docking. *ChemComm* **2013,** *49* (10), 981–983.

[273] Hassel, O. Structural Aspects of Interatomic Charge-Transfer Bonding. *Science* **1970,** *170* (3957), 497–502.

[274] Vulpetti, A.; Dalvit, C. Fluorine Local Environment: from Screening to Drug Design. *Drug Discovery Today* **2012,** *17* (15–16), 890–897.

[275] Dalvit, C.; Vulpetti, A. Intermolecular and Intramolecular Hydrogen Bonds Involving Fluorine Atoms: Implications for Recognition, Selectivity, and Chemical Properties. *ChemMedChem* **2012,** *7* (2), 262–272.

[276] Müller, K.; Faeh, C.; Diederich, F. Fluorine in Pharmaceuticals: Looking Beyond Intuition. *Science* **2007,** *317* (5846), 1881–1886.

[277] Harding, M. M. The Geometry of Metal-Ligand Interactions Relevant to Proteins. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1999,** *55* (8), 1432–1443.

[278] Harding, M. M. Geometry of Metal-Ligand Interactions in Proteins. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2001,** *57* (3), 401–411.

[279] Jacobsen, F. E.; Lewis, J. A.; Cohen, S. M. The Design of Inhibitors for Medicinally Relevant Metalloproteins. *ChemMedChem* **2007,** *2* (2), 152–171.

[280] Seebeck, B.; Reulecke, I.; Kämper, A.; Rarey, M. Modeling of Metal Interaction Geometries for Protein–Ligand Docking. *Proteins: Struct., Funct., Bioinf.* **2008,** *71* (3), 1237–1254.

[281] Hu, X.; Balaz, S.; Shelver, W. H. A Practical Approach to Docking of Zinc Metalloproteinase Inhibitors. *J. Mol. Graphics Modell.* **2004,** *22* (4), 293–307.

[282] Irwin, J. J.; Raushel, F. M.; Shoichet, B. K. Virtual Screening against Metalloenzymes for Inhibitors and Substrates. *Biochemistry* **2005,** *44* (37), 12316–12328.

[283] Langer, T.; Wolber, G. Pharmacophore Definition and 3D Searches. *Drug Discovery Today: Technol.* **2004,** *1* (3), 203–207.

[284] Güner, O. F., Ed. *Pharmacophore: Perception, Development, and Use in Drug Design;* International University Line: La Jolla, CA, USA, 2000; IUL Biotechnology Serie, Bd. 2.

[285] Langer, T., Hoffmann, R. D., Eds. *Pharmacophores and Pharmacophore Searches;* Wiley-VCH: Weinheim, 2006; Methods and Principles in Medicinal Chemistry, Bd. 32.

[286] Leach, A. R.; Gillet, V. J.; Lewis, R. A.; Taylor, R. Three-Dimensional Pharmacophore Methods in Drug Discovery. *J. Med. Chem.* **2010,** *53* (2), 539–558.

[287] Greene, J.; Kahn, S.; Savoj, H.; Sprague, P.; Teig, S. Chemical Function Queries for 3D Database Search. *J. Chem. Inf. Comput. Sci.* **1994,** *34* (6), 1297–1308.

[288] Catalyst. Accelrys, Accelrys Inc. *www.accelrys.com.*

[289] Taminau, J.; Thijs, G.; De Winter, H. Pharao: Pharmacophore Alignment and Optimization. *J. Mol. Graphics Modell.* **2008,** *27* (2), 161–169.

[290] Wolber, G.; Seidel, T.; Bendix, F.; Langer, T. Molecule-Pharmacophore Superpositioning and Pattern Matching in Computational Drug Design. *Drug Discovery Today* **2008,** *13* (1–2), 23–29.

[291] Zuccotto, F. Pharmacophore Features Distributions in Different Classes of Compounds. *J. Chem. Inf. Comput. Sci.* **2003,** *43* (5), 1542–1552.

[292] Kubinyi, H. Drug Research: Myths, Hype and Reality. *Nat. Rev. Drug Discovery* **2003,** *2* (8), 665–668.

[293] Oellien, F.; Cramer, J.; Beyer, C.; Ihlenfeldt, W.-D.; Selzer, P. M. The Impact of Tautomer Forms on Pharmacophore-Based Virtual Screening†. *J. Chem. Inf. Model.* **2006,** *46* (6), 2342–2354.

[294] Martin, Y. Let's Not Forget Tautomers. *J. Comput.-Aided Mol. Des.* **2009,** *23* (10), 693-704.

[295] Böhm, H.-J.; Brode, S.; Hesse, U.; Klebe, G. Oxygen and Nitrogen in Competitive Situations: Which is the Hydrogen-Bond Acceptor? *Chem. Eur. J.* **1996,** *2* (12), 1509–1513.

[296] Pierce, A. C.; Sandretto, K. L.; Bemis, G. W. Kinase Inhibitors and the Case for CH…O Hydrogen Bonds in Protein–Ligand Binding. *Proteins: Struct., Funct., Bioinf.* **2002,** *49* (4), 567–576.

[297] Spitzer, G. M.; Heiss, M.; Mangold, M.; Markt, P.; Kirchmair, J.; Wolber, G.; Liedl, K. R. One Concept, Three Implementations of 3D Pharmacophore-Based Virtual Screening: Distinct Coverage of Chemical Search Space. *J. Chem. Inf. Model.* **2010,** *50* (7), 1241–1247.

[298] Dixon, S.; Smondyrev, A.; Knoll, E.; Rao, S.; Shaw, D.; Friesner, R. PHASE: A New Engine for Pharmacophore Perception, 3D QSAR Model Development, and 3D Database Screening: 1. Methodology and Preliminary Results. *J. Comput.-Aided Mol. Des.* **2006,** *20* (10-11), 647-671.

[299] Wolber, G.; Langer, T. LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters. *J. Chem. Inf. Model.* **2005,** *45* (1), 160–169.

[300] Kossner, M. T. *Pharmacophoric Distance Profiles (DIP^2) for Virtual Screening (Dissertation);* Technische Universität Braunschweig: Braunschweig, 2011.

[301] Todorov, N. P.; Alberts, I. L.; de Esch, Iwan J. P.; Dean, P. M. QUASI: A Novel Method for Simultaneous Superposition of Multiple Flexible Ligands and Virtual Screening Using Partial Similarity. *J. Chem. Inf. Model.* **2007,** *47* (3), 1007–1020.

[302] Lloyd, D. What Is Aromaticity? *J. Chem. Inf. Comput. Sci.* **1996,** *36* (3), 442–447.

[303] Roos-Kozel, B. L.; Jorgensen, W. L. Computer-Assisted Mechanistic Evaluation of Organic Reactions. 2. Perception of Rings, Aromaticity, and Tautomers. *J. Chem. Inf. Comput. Sci.* **1981,** *21* (2), 101–111.

[304] *OpenEye Scientific Software. Handbuch OEChem TK - Python Release 1.9.2 (vom 13.06.2013). www.eyesopen.com/docs/toolkits/current/pdf/OEChem_TK-python.pdf*.

[305] Wermuth, C. G. Multitargeted Drugs: The End of the 'One-Target-One-Disease' Philosophy? *Drug Discovery Today* **2004,** *9* (19), 826–827.

[306] Reddy, A. S.; Zhang, S. Polypharmacology: Drug Discovery for the Future. *Expert Rev. Clin. Pharmacol.* **2012,** *6* (1), 41–47.

[307] Peters, J.-U. Polypharmacology – Foe or Friend? *J. Med. Chem.* **2013,** *56* (22), 8955–8971.

[308] Vaz, R. J.; Klabunde, T. *Antitargets*. *Prediction and Prevention of Drug Side Effects*; Methods and Principles in Medicinal Chemistry, Bd. 38; Wiley-VCH: Weinheim, 2008.

[309] Hamon, J.; Whitebread, S. In Vitro Safety Pharmacology Profiling: An Important Tool to Decrease Attrition. In *Hit and Lead Profiling*: *Identification and Optimization of Drug-Like Molecules*; Faller, B., Urban, L., Eds.; Wiley-VCH: Weinheim, 2010, Bd. 43; S. 273–295.

[310] Whitebread, S.; Hamon, J.; Bojanic, D.; Urban, L. Keynote Review: In Vitro Safety Pharmacology Profiling: An Essential Tool for Successful Drug Development. *Drug Discovery Today* **2005,** *10* (21), 1421–1433.

[311] Bass, A. S.; Cartwright, M. E.; Mahon, C.; Morrison, R.; Snyder, R.; McNamara, P.; Bradley, P.; Zhou, Y.-Y.; Hunter, J. Exploratory Drug Safety: A Discovery Strategy to Reduce Attrition in Development. *J. Pharmacol. Toxicol. Methods* **2009,** *60* (1), 69–78.

[312] Raju, T. N. The Nobel Chronicles. 1988: James Whyte Black,(b 1924), Gertrude Elion (1918-99), and George H Hitchings (1905-98). *Lancet* **2000,** *355* (9208), 1022.

[313] Wermuth, C. G. Selective Optimization of Side Activities: Another Way for Drug Discovery. *J. Med. Chem.* **2004,** *47* (6), 1303–1314.

[314] Caron, P. R.; Mullican, M. D.; Mashal, R. D.; Wilson, K. P.; Su, M. S.; Murcko, M. A. Chemogenomic Approaches to Drug Discovery. *Curr. Opin. Chem. Biol.* **2001,** *5* (4), 464–470.

[315] Rognan, D. Chemogenomic Approaches to Rational Drug Design. *Br. J. Pharmacol.* **2007,** *152* (1), 38–52.

[316] Klabunde, T. Chemogenomic Approaches to Drug Discovery: Similar Receptors Bind Similar Ligands. *Br. J. Pharmacol.* **2007,** *152* (1), 5–7.

[317] Kubinyi, H. Chemogenomics in Drug Discovery. In *Chemical Genomics*; Jaroch, S., Weinmann, H., Eds.; Springer: Berlin, 2006; Ernst Schering Research Foundation Workshop, Bd. 58; S. 1-19.

[318] Bredel, M.; Jacoby, E. Chemogenomics: An Emerging Strategy for Rapid Target and Drug Discovery. *Nat. Rev. Genet.* **2004,** *5* (4), 262–275.

[319] Kubinyi, H., Müller, G., Eds. *Chemogenomics in Drug Discovery;* Wiley-VCH: Weinheim, 2004; Methods and Principles in Medicinal Chemistry, Bd. 22.

[320] Hu, Y.; Bajorath, J. Compound Promiscuity: What Can We Learn from Current Data? *Drug Discovery Today* **2013,** *18* (13–14), 644–650.

[321] Evans, B. E.; Rittle, K. E.; Bock, M. G.; DiPardo, R. M.; Freidinger, R. M.; Whitter, W. L.; Lundell, G. F.; Veber, D. F.; Anderson, P. S. Methods for Drug Discovery: Development of Potent, Selective, Orally Effective Cholecystokinin Antagonists. *J. Med. Chem.* **1988,** *31* (12), 2235–2246.

[322] DeSimone, R. W.; Currie, K. S.; Mitchell, S. A.; Darrow, J. W.; Pippin, D. A. Privileged Structures: Applications in Drug Discovery. *Comb. Chem. High Throughput Screening* **2004,** *7* (5), 473–493.

[323] Patchett, A. A.; Nargund, R. P. Chapter 26. Privileged Structures - An Update. In *Annual Reports in Medicinal Chemistry*, Volume 35; Doherty, A. M., Greenlee, W. F., Eds.; Academic Press: San Diego, 2000, Bd. 35; S. 289–298.

[324] Achenbach, J.; Tiikkainen, P.; Franke, L.; Proschak, E. Computational Tools for Polypharmacology and Repurposing. *Future Med. Chem.* **2011,** *3* (8), 961–968.

[325] Hu, Y.; Stumpfe, D.; Bajorath, J. Visualization of Activity Landscapes and Chemogenomics Data. *Mol. Inf.* **2013,** *32* (11-12), 954–963.

[326] Jenkins, J. L.; Bender, A.; Davies, J. W. In Silico Target Fishing: Predicting Biological Targets from Chemical Structure. *Drug Discovery Today: Technol.* **2006,** *3* (4), 413–421.

[327] Bender, A.; Young, D. W.; Jenkins, J. L.; Serrano, M.; Mikhailov, D.; Clemons, P. A.; Davies, J. W. Chemogenomic Data Analysis: Prediction of Small-Molecule Targets and the Advent of Biological Fingerprints. *Comb. Chem. High Throughput Screening* **2007,** *10* (8), 719–731.

[328] Scheiber, J.; Bender, A.; Azzaoui, K.; Jenkins, J. Knowledge-Based and Computational Approaches to In Vitro Safety Pharmacology. In *Hit and Lead Profiling*: *Identification and Optimization of Drug-Like Molecules*; Faller, B., Urban, L., Eds.; Wiley-VCH: Weinheim, 2010, Bd. 43; S. 297–322.

[329] Hopkins, A. L. Network Pharmacology: The Next Paradigm in Drug Discovery. *Nat. Chem. Biol.* **2008,** *4* (11), 682–690.

[330] Mestres, J.; Gregori-Puigjane, E.; Valverde, S.; Sole, R. V. Data Completeness - The Achilles Heel of Drug-Target Networks. *Nat. Biotechnol.* **2008,** *26* (9), 983–984.

[331] Weber, A.; Casini, A.; Heine, A.; Kuhn, D.; Supuran, C. T.; Scozzafava, A.; Klebe, G. Unexpected Nanomolar Inhibition of Carbonic Anhydrase by COX-2-Selective Celecoxib: New Pharmacological Opportunities Due to Related Binding Site Recognition. *J. Med. Chem.* **2004,** *47* (3), 550–557.

[332] Frimurer, T. M.; Ulven, T.; Elling, C. E.; Gerlach, L.-O.; Kostenis, E.; Högberg, T. A Physicogenetic Method to Assign Ligand-Binding Relationships between 7TM Receptors. *Bioorg. Med. Chem. Lett.* **2005,** *15* (16), 3707–3712.

[333] Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990,** *215* (3), 403–410.

[334] McGinnis, S.; Madden, T. L. BLAST: At the Core of a Powerful and Diverse Set of Sequence Analysis Tools. *Nucleic Acids Res.* **2004,** *32* (suppl 2), W20-W25.

[335] Surgand, J.-S.; Rodrigo, J.; Kellenberger, E.; Rognan, D. A Chemogenomic Analysis of the Transmembrane Binding Cavity of Human G-Protein-Coupled Receptors. *Proteins: Struct., Funct., Bioinf.* **2006,** *62* (2), 509–538.

[336] Manning, G. The Protein Kinase Complement of the Human Genome. *Science* **2002,** *298* (5600), 1912–1934.

[337] Rawlings, N. D.; Morton, F. R.; Kok, C. Y.; Kong, J.; Barrett, A. J. MEROPS: The Peptidase Database. *Nucleic Acids Res.* **2008,** *36* (suppl 1), D320-D325.

[338] Rawlings, N. D.; Barrett, A. J.; Bateman, A. MEROPS: The Database of Proteolytic Enzymes, Their Substrates and Inhibitors. *Nucleic Acids Res.* **2012,** *40* (D1), D343-D350.

[339] Kuhn, D.; Weskamp, N.; Schmitt, S.; Hüllermeier, E.; Klebe, G. From the Similarity Analysis of Protein Cavities to the Functional Classification of Protein Families Using Cavbase. *J. Mol. Biol.* **2006,** *359* (4), 1023–1044.

[340] Schmitt, S.; Kuhn, D.; Klebe, G. A New Method to Detect Related Function Among Proteins Independent of Sequence and Fold Homology. *J. Mol. Biol.* **2002,** *323* (2), 387–406.

[341] Vulpetti, A.; Kalliokoski, T.; Milletti, F. Chemogenomics in Drug Discovery: Computational Methods Based on the Comparison of Binding Sites. *Future Med. Chem.* **2012,** *4* (15), 1971–1979.

[342] Paolini, G. V.; Shapland, Richard H B; van Hoorn, Willem P; Mason, J. S.; Hopkins, A. L. Global Mapping of Pharmacological Space. *Nat. Biotechnol.* **2006,** *24* (7), 805–815.

[343] van der Horst, E.; Peironcely, J.; IJzerman, A.; Beukers, M.; Lane, J.; van Vlijmen, H.; Emmerich, M.; Okuno, Y.; Bender, A. A Novel Chemogenomics Analysis of G Protein-Coupled

Receptors (GPCRs) and Their Ligands: A Potential Strategy for Receptor De-Orphanization. *BMC Bioinf.* **2010,** *11* (1), 316.

[344] Sutherland, J. J.; Higgs, R. E.; Watson, I.; Vieth, M. Chemical Fragments as Foundations for Understanding Target Space and Activity Prediction. *J. Med. Chem.* **2008,** *51* (9), 2689–2700.

[345] Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating Protein Pharmacology by Ligand Chemistry. *Nat. Biotechnol.* **2007,** *25* (2), 197–206.

[346] Hert, J.; Keiser, M. J.; Irwin, J. J.; Oprea, T. I.; Shoichet, B. K. Quantifying the Relationships among Drug Classes. *J. Chem. Inf. Model.* **2008,** *48* (4), 755–765.

[347] Keiser, M. J.; Setola, V.; Irwin, J. J.; Laggner, C.; Abbas, A. I.; Hufeisen, S. J.; Jensen, N. H.; Kuijer, M. B.; Matos, R. C.; Tran, T. B.; Whaley, R.; Glennon, R. A.; Hert, J.; Thomas, Kelan L. H.; Edwards, D. D.; Shoichet, B. K.; Roth, B. L. Predicting New Molecular Targets for Known Drugs. *Nature* **2009,** *462* (7270), 175–181.

[348] Lounkine, E.; Keiser, M. J.; Whitebread, S.; Mikhailov, D.; Hamon, J.; Jenkins, J. L.; Lavan, P.; Weber, E.; Doak, A. K.; Cote, S.; Shoichet, B. K.; Urban, L. Large-Scale Prediction and Testing of Drug Activity on Side-Effect Targets. *Nature* **2012,** *486* (7403), 361–367.

[349] Lin, H.; Sassano, M. F.; Roth, B. L.; Shoichet, B. K. A Pharmacological Organization of G Protein-Coupled Receptors. *Nat. Methods* **2013,** *10* (2), 140–146.

[350] Gregori-Puigjané, E.; Mestres, J. SHED: Shannon Entropy Descriptors from Topological Feature Distributions. *J. Chem. Inf. Model.* **2006,** *46* (4), 1615–1622.

[351] Gregori-Puigjané, E.; Mestres, J. A Ligand-Based Approach to Mining the Chemogenomic Space of Drugs. *Comb. Chem. High Throughput Screening* **2008,** *11* (8), 669–676.

[352] Mestres, J.; Martín-Couce, L.; Gregori-Puigjané, E.; Cases, M.; Boyer, S. Ligand-Based Approach to In Silico Pharmacology: Nuclear Receptor Profiling. *J. Chem. Inf. Model.* **2006,** *46* (6), 2725–2736.

[353] Nidhi; Glick, M.; Davies, J. W.; Jenkins, J. L. Prediction of Biological Targets for Compounds Using Multiple-Category Bayesian Models Trained on Chemogenomics Databases. *J. Chem. Inf. Model.* **2006,** *46* (3), 1124–1133.

[354] Nigsch, F.; Bender, A.; Jenkins, J. L.; Mitchell, John B. O. Ligand-Target Prediction Using Winnow and Naive Bayesian Algorithms and the Implications of Overall Performance Statistics. *J. Chem. Inf. Model.* **2008,** *48* (12), 2313–2325.

[355] Koutsoukas, A.; Lowe, R.; KalantarMotamedi, Y.; Mussa, H. Y.; Klaffke, W.; Mitchell, John B. O.; Glen, R. C.; Bender, A. In Silico Target Predictions: Defining a Benchmarking Data Set and Comparison of Performance of the Multiclass Naïve Bayes and Parzen-Rosenblatt Window. *J. Chem. Inf. Model.* **2013,** *53* (8), 1957–1966.

[356] Grant, J. A.; Gallardo, M. A.; Pickup, B. T. A Fast Method of Molecular Shape Comparison: A Simple Application of a Gaussian Description of Molecular Shape. *J. Comput. Chem.* **1996,** *17* (14), 1653–1666.

[357] Rush, T. S.; Grant, J. A.; Mosyak, L.; Nicholls, A. A Shape-Based 3-D Scaffold Hopping Method and Its Application to a Bacterial Protein-Protein Interaction. *J. Med. Chem.* **2005,** *48* (5), 1489–1495.

[358] AbdulHameed, M. D. M.; Chaudhury, S.; Singh, N.; Sun, H.; Wallqvist, A.; Tawa, G. J. Exploring Polypharmacology Using a ROCS-Based Target Fishing Approach. *J. Chem. Inf. Model.* **2011,** *52* (2), 492–505.

[359] Vasudevan, S. R.; Moore, J. B.; Schymura, Y.; Churchill, G. C. Shape-Based Reprofiling of FDA-Approved Drugs for the H1 Histamine Receptor. *J. Med. Chem.* **2012,** *55* (16), 7054–7060.

[360] Liu, X.; Ouyang, S.; Yu, B.; Liu, Y.; Huang, K.; Gong, J.; Zheng, S.; Li, Z.; Li, H.; Jiang, H. PharmMapper Server: A Web Server for Potential Drug Target Identification Using Pharmacophore Mapping Approach. *Nucleic Acids Res.* **2010,** *38* (suppl 2), W609-W614.

[361] Jenkins, J. L.; Glick, M.; Davies, J. W. A 3D Similarity Method for Scaffold Hopping from Known Drugs or Natural Ligands to New Chemotypes. *J. Med. Chem.* **2004,** *47* (25), 6144–6159.

[362] Nettles, J. H.; Jenkins, J. L.; Bender, A.; Deng, Z.; Davies, J. W.; Glick, M. Bridging Chemical and Biological Space: "Target Fishing" Using 2D and 3D Molecular Descriptors. *J. Med. Chem.* **2006,** *49* (23), 6802–6810.

[363] Li, H.; Gao, Z.; Kang, L.; Zhang, H.; Yang, K.; Yu, K.; Luo, X.; Zhu, W.; Chen, K.; Shen, J.; Wang, X.; Jiang, H. TarFisDock: A Web Server for Identifying Drug Targets with Docking Approach. *Nucleic Acids Res.* **2006,** *34* (suppl 2), W219-W224.

[364] Chen, Y. Z.; Zhi, D. G. Ligand–Protein Inverse Docking and Its Potential Use in the Computer Search of Protein Targets of a Small Molecule. *Proteins: Struct., Funct., Bioinf.* **2001,** *43* (2), 217–226.

[365] Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nat. Rev. Drug Discovery* **2004,** *3* (11), 935–949.

[366] Scheiber, J.; Chen, B.; Milik, M.; Sukuru, Sai Chetan K.; Bender, A.; Mikhailov, D.; Whitebread, S.; Hamon, J.; Azzaoui, K.; Urban, L.; Glick, M.; Davies, J. W.; Jenkins, J. L. Gaining Insight into Off-Target Mediated Effects of Drug Candidates with a Comprehensive Systems Chemical Biology Analysis. *J. Chem. Inf. Model.* **2009,** *49* (2), 308–317.

[367] Scheiber, J.; Jenkins, J. L.; Sukuru, Sai Chetan K.; Bender, A.; Mikhailov, D.; Milik, M.; Azzaoui, K.; Whitebread, S.; Hamon, J.; Urban, L.; Glick, M.; Davies, J. W. Mapping Adverse Drug Reactions in Chemical Space. *J. Med. Chem.* **2009,** *52* (9), 3103–3107.

[368] Bender, A.; Scheiber, J.; Glick, M.; Davies, J. W.; Azzaoui, K.; Hamon, J.; Urban, L.; Whitebread, S.; Jenkins, J. L. Analysis of Pharmacology Data and the Prediction of Adverse Drug Reactions and Off-Target Effects from Chemical Structure. *ChemMedChem* **2007,** *2* (6), 861–873.

[369] Daylight Chemical Information Systems. SMARTS - A Language for Describing Molecular Patterns. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.

[370] OEChemTK 1.9.0. OpenEye Scientific Software Inc., Santa Fe, NM. *http://www.eyesopen.com/.*

[371] Figueras, J. Ring Perception Using Breadth-First Search. *J. Chem. Inf. Comput. Sci.* **1996,** *36* (5), 986–991.

[372] Berger, F.; Flamm, C.; Gleiss, P. M.; Leydold, J.; Stadler, P. F. Counterexamples in Chemical Ring Perception. *J. Chem. Inf. Comput. Sci.* **2004,** *44* (2), 323–331.

[373] Kolodzik, A.; Urbaczek, S.; Rarey, M. Unique Ring Families: A Chemically Meaningful Description of Molecular Ring Topologies. *J. Chem. Inf. Model.* **2012,** *52* (8), 2013–2021.

[374] Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **1996,** *17* (5-6), 490–519.

[375] NetworkX 1.6. *http://networkx.lanl.gov/.*

[376] Hagberg, A. A.; Schult, D. A.; Swart, P. J. Exploring Network Structure, Dynamics, and Function Using NetworkX. In *Proceedings of the 7th Python in Science Conference*; Varoquaux, G., Vaught, T., Millman, J., Eds.; SciPy2008: Pasadena, CA, USA, 2008; S. 11–15.

[377] Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003,** *13* (11), 2498–2504.

[378] Cline, M. S.; Smoot, M.; Cerami, E.; Kuchinsky, A.; Landys, N.; Workman, C.; Christmas, R.; Avila-Campilo, I.; Creech, M.; Gross, B.; Hanspers, K.; Isserlin, R.; Kelley, R.; Killcoyne, S.; Lotia, S.; Maere, S.; Morris, J.; Ono, K.; Pavlovic, V.; Pico, A. R.; Vailaya, A.; Wang, P.-L.; Adler, A.; Conklin, B. R.; Hood, L.; Kuiper, M.; Sander, C.; Schmulevich, I.; Schwikowski, B.; Warner, G. J.; Ideker, T.; Bader, G. D. Integration of Biological Networks and Gene Expression Data Using Cytoscape. *Nat. Protoc.* **2007,** *2* (10), 2366–2382.

[379] Smoot, M. E.; Ono, K.; Ruscheinski, J.; Wang, P.-L.; Ideker, T. Cytoscape 2.8: New Neatures for Data Integration and Network Visualization. *Bioinformatics* **2011,** *27* (3), 431–432.

[380] Cytoscape 2.8.2. *http://www.cytoscape.org/.*

[381] Fruchterman, T. M. J.; Reingold, E. M. Graph Drawing by Force-Directed Placement. *Softw. Pract. Exper.* **1991,** *21* (11), 1129–1164.

[382] Wallace, I. M.; Bader, G. D.; Giaever, G.; Nislow, C. Displaying Chemical Information on a Biological Network Using Cytoscape. *Methods Mol. Biol.* **2011,** *781*, 363–376.

[383] Chemoinformatics Plugin for Cytoscape (UCSF chemViz). *http://www.cgl.ucsf.edu/ cytoscape/chemViz/.*

[384] Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. A Comparison of Methods for Modeling Quantitative Structure−Activity Relationships. *J. Med. Chem.* **2004,** *47* (22), 5541–5554.

[385] Fontaine, F.; Pastor, M.; Zamora, I.; Sanz, F. Anchor−GRIND: Filling the Gap between Standard 3D QSAR and the GRid-INdependent Descriptors. *J. Med. Chem.* **2005,** *48* (7), 2687–2694.

[386] Chen, X.; Reynolds, C. H. Performance of Similarity Measures in 2D Fragment-Based Similarity Searching: Comparison of Structural Descriptors and Similarity Coefficients. *J. Chem. Inf. Comput. Sci.* **2002,** *42* (6), 1407–1414.

[387] Lipinski, C. A. Drug-Like Properties and the Causes of Poor Solubility and Poor Permeability. *J. Pharmacol. Toxicol. Methods* **2000,** *44* (1), 235–249.

[388] Oprea, T. I. Property Distribution of Drug-Related Chemical Databases. *J. Comput.-Aided Mol. Des.* **2000,** *14* (3), 251-264.

[389] Wassermann, A. M.; Bajorath, J. Identification of Target Family Directed Bioisosteric Replacements. *MedChemComm* **2011,** *2* (7), 601–606.

[390] Irwin, J. J.; Shoichet, B. K. ZINC − A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005,** *45* (1), 177–182.

[391] Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* **2012,** *52* (7), 1757–1768.

[392] ZINC12. *http://zinc.docking.org/.*

[393] Spearman, C. The Proof and Measurement of Association between Two Things. *Am. J. Psychol.* **1904,** *15* (1), 72–101.

[394] Pearson, K. Notes on the History of Correlation. *Biometrika* **1920,** *13* (1), 25–45.

[395] Artusi, R.; Verderio, P.; Marubini, E. Bravais-Pearson and Spearman Correlation Coefficients: Meaning, Test of Hypothesis and Confidence Interval. *Int. J. Biol. Markers* **2002,** *17*, 148–151.

[396] Kumar, S. Semantic Clustering of Index Terms. *J. ACM* **1968,** *15* (4), 493–513.

[397] Jain, A. K.; Murty, M. N.; Flynn, P. J. Data Clustering: A Review. *ACM Comput. Surv.* **1999,** *31* (3), 264–323.

[398] Jain, A. K. Data Clustering: 50 Years Beyond K-Means. *Pattern Recogn. Lett.* **2010,** *31* (8), 651–666.

[399] Hertz, J. A.; Krogh, A. S.; Palmer, R. G. *Introduction to the Theory of Neural Computation;* Addison-Wesley: Redwood City, CA, 1991.

[400] Kohonen, T. *Self-Organizing Maps*, 3. Aufl.; Springer: Berlin, 2001.

[401] Tibshirani, R.; Walther, G.; Hastie, T. Estimating the Number of Clusters in a Data Set via the Gap Statistic. *J. R. Stat. Soc. Series B Stat. Methodol.* **2001,** *63* (2), 411–423.

[402] SVL-Script zur Berechnung vom ECFP4 Fingerprint (MOE-SVL-Exchange): ph4_ExtendedConnectivityFP.svl (Autor: Markus Kossner, Version: 2012-12-10). *http://svl.chemcomp.com/.*

[403] Duda, R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification*, 2. Aufl.; John Wiley & Sons: New York, 2001.

[404] Altman, N. S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* **1992,** *46* (3), 175–185.

[405] Zheng, W.; Tropsha, A. Novel Variable Selection Quantitative Structure−Property Relationship Approach Based on the k-Nearest-Neighbor Principle. *J. Chem. Inf. Comput. Sci.* **1999,** *40* (1), 185–194.

[406] Shen, M.; LeTiran, A.; Xiao, Y.; Golbraikh, A.; Kohn, H.; Tropsha, A. Quantitative Structure−Activity Relationship Analysis of Functionalized Amino Acid Anticonvulsant Agents Using k Nearest Neighbor and Simulated Annealing PLS Methods. *J. Med. Chem.* **2002,** *45* (13), 2811–2823.

[407] SVL-Script zur Berechnung des MACCSF-Fingerprints (MOE-SVL-Exchange): ph4-maccsf.svl (in "new-fingerprints.zip", Author: Todd Ewing, Version: 2007-01-12). *http://svl.chemcomp.com/.*

[408] Python-Skript zur Berechnung des CATS2D-Fingerprints (cats2d.py, Autor: R. Guha, Version: 26.08.2007). *http://rghuha.net.*

[409] Kessler, W. *Multivariate Datenanalyse: für die Pharma, Bio- und Prozessanalytik;* Wiley-VCH: Weinheim, 2007; S. 89–152.

[410] Medina-Franco, J. L.; Edwards, B. S.; Pinilla, C.; Appel, J. R.; Giulianotti, M. A.; Santos, R. G.; Yongye, A. B.; Sklar, L. A.; Houghten, R. A. Rapid Scanning Structure–Activity Relationships in Combinatorial Data Sets: Identification of Activity Switches. *J. Chem. Inf. Model.* **2013,** *53* (6), 1475–1485.

[411] Cases, M.; Mestres, J. A Chemogenomic Approach to Drug Discovery: Focus on Cardiovascular Diseases. *Drug Discovery Today* **2009,** *14* (9-10), 479–485.

[412] Klabunde, T.; Hessler, G. Drug Design Strategies for Targeting G-Protein-Coupled Receptors. *ChemBioChem* **2002,** *3* (10), 928–944.

[413] Kaczorowski, G. J.; McManus, O. B.; Priest, B. T.; Garcia, M. L. Ion Channels as Drug Targets: The Next GPCRs. *J. Gen. Physiol.* **2008,** *131* (5), 399–405.

[414] Müller, G. Target Family-Directed Masterkeys in Chemogenomics. In *Chemogenomics in Drug Discovery*; Kubinyi, H., Müller, G., Eds.; Wiley-VCH: Weinheim, 2004; Methods and Principles in Medicinal Chemistry, Bd. 22; S. 5–41.

[415] Gloriam, D. E.; Foord, S. M.; Blaney, F. E.; Garland, S. L. Definition of the G Protein-Coupled Receptor Transmembrane Bundle Binding Pocket and Calculation of Receptor Similarities for Drug Design. *J. Med. Chem.* **2009,** *52* (14), 4429–4442.

[416] Cell Signaling Technology; Manning, G. The Human Protein Kinases. *http://www.cellsignal.com/reference/kinase/.*

[417] MEROPS. The Peptidase Database (Release 9.9). *http://merops.sanger.ac.uk/.*

[418] A Unified Nomenclature System for the Nuclear Receptor Superfamily. *Cell* **1999,** *97* (2), 161–163.

[419] International Union of Biochemistry and Molecular Biology. Recommendations on Biochemical & Organic Nomenclature, Symbols & Terminology. *http://www.chem.qmul.ac.uk/iubmb/.*

[420] Bayer Pharma AG (Berlin). Fachinformationen Xarelto$^{(R)}$ 10mg/15mg/20mg Filmtabletten (Stand: Juni 2013). *http://www.fachinfo.de/*.

[421] Bristol-Myers Squibb/Pfizer EEIG (Uxbridge, UK). Fachinformationen Eliquis$^{(R)}$ 2,5mg/5mg Filmtabletten (Stand: September 2013). *http://www.fachinfo.de/*.

[422] Pinto, D. J. P.; Smallheer, J. M.; Cheney, D. L.; Knabb, R. M.; Wexler, R. R. Factor Xa Inhibitors: Next-Generation Antithrombotic Agents. *J. Med. Chem.* **2010,** *53* (17), 6243–6274.

[423] Schechter, I.; Berger, A. On the Size of the Active Site in Proteases. I. Papain. *Biochem. Biophys. Res. Commun.* **1967,** *27* (2), 157–162.

[424] Mochizuki, A.; Nagata, T.; Kanno, H.; Takano, D.; Kishida, M.; Suzuki, M.; Ohta, T. Orally Active Zwitterionic Factor Xa Inhibitors with Long Duration of Action. *Bioorg. Med. Chem. Lett.* **2011,** *21* (24), 7337–7343.

[425] Zhu, B.-Y.; Jia, Z. J.; Zhang, P.; Su, T.; Huang, W.; Goldman, E.; Tumas, D.; Kadambi, V.; Eddy, P.; Sinha, U.; Scarborough, R. M.; Song, Y. Inhibitory Effect of Carboxylic Acid Group on hERG Binding. *Bioorg. Med. Chem. Lett.* **2006,** *16* (21), 5507–5512.

[426] Sanguinetti, M. C.; Tristani-Firouzi, M. hERG Potassium Channels and Cardiac Arrhythmia. *Nature* **2006,** *440* (7083), 463–469.

[427] Ekroos, M.; Sjögren, T. Structural Basis for Ligand Promiscuity in Cytochrome P450 3A4. *Proc. Natl. Acad. Sci. U. S. A.* **2006,** *103* (37), 13682–13687.

[428] Aspiotis, R.; Chen, A.; Cauchon, E.; Dubé, D.; Falgueyret, J.-P.; Gagné, S.; Gallant, M.; Grimm, E. L.; Houle, R.; Juteau, H.; Lacombe, P.; Laliberté, S.; Lévesque, J.-F.; MacDonald, D.; McKay, D.; Percival, M. David; Roy, P.; Soisson, S. M.; Wu, T. The Discovery and Synthesis of Potent Zwitterionic Inhibitors of Renin. *Bioorg. Med. Chem. Lett.* **2011,** *21* (8), 2430–2436.

[429] Su, T.; Wu, Y.; Doughan, B.; Kane-Maguire, K.; Marlowe, C. K.; Kanter, J. P.; Woolfrey, J.; Huang, B.; Wong, P.; Sinha, U.; Park, G.; Malinowski, J.; Hollenbach, S.; Scarborough, R. M.; Zhu, B.-Y. Design and Synthesis of Glycolic and Mandelic Acid Derivatives as Factor Xa Inhibitors. *Bioorg. Med. Chem. Lett.* **2001,** *11* (17), 2279–2282.

[430] Song, Y.; Clizbe, L.; Bhakta, C.; Teng, W.; Li, W.; Wu, Y.; Jia, Z. J.; Zhang, P.; Wang, L.; Doughan, B.; Su, T.; Kanter, J.; Woolfrey, J.; Wong, P.; Huang, B.; Tran, K.; Sinha, U.; Park, G.; Reed, A.; Malinowski, J.; Hollenbach, S.; Scarborough, R. M.; Zhu, B.-Y. Design, Synthesis, and SAR of Substituted Acrylamides as Factor Xa Inhibitors. *Bioorg. Med. Chem. Lett.* **2002,** *12* (11), 1511–1515.

[431] Shi, Y.; Sitkoff, D.; Zhang, J.; Han, W.; Hu, Z.; Stein, P. D.; Wang, Y.; Kennedy, L. J.; O'Connor, S. P.; Ahmad, S.; Liu, Eddie C. -K.; Seiler, S. M.; Lam, Patrick Y. S.; Robl, J. A.; Macor, J. E.; Atwal, K. S.; Zahler, R. Amino(methyl) Pyrrolidines as Novel Scaffolds for Factor Xa Inhibitors. *Bioorg. Med. Chem. Lett.* **2007,** *17* (21), 5952–5958.

[432] Roehrig, S.; Straub, A.; Pohlmann, J.; Lampe, T.; Pernerstorfer, J.; Schlemmer, K.-H.; Reinemer, P.; Perzborn, E. Discovery of the Novel Antithrombotic Agent 5-Chloro-N-({(5S)-2-oxo-3- [4-(3-oxomorpholin-4-yl)phenyl]-1,3-oxazolidin-5-yl}methyl)thiophene- 2-carboxamide (BAY 59-7939): An Oral, Direct Factor Xa Inhibitor. *J. Med. Chem.* **2005,** *48* (19), 5900–5908.

[433] Straub, A.; Roehrig, S.; Hillisch, A. Entering the Era of Non-Basic P1 Site Groups: Discovery of Xarelto (Rivaroxaban). *Curr. Top. Med. Chem.* **2010,** *10* (3), 257–269.

[434] Knockaert, M.; Greengard, P.; Meijer, L. Pharmacological Inhibitors of Cyclin-Dependent Kinases. *Trends Pharmacol. Sci.* **2002,** *23* (9), 417–425.

[435] Hirama, T.; Koeffler, H. P. Role of the Cyclin-Dependent Kinase Inhibitors in the Development of Cancer. *Blood* **1995,** *86* (3), 841–854.

[436] Noble, M. E. M.; Endicott, J. A.; Johnson, L. N. Protein Kinase Inhibitors: Insights into Drug Design from Structure. *Science* **2004,** *303* (5665), 1800–1805.

[437] Schulze-Gahmen, U.; De Bondt, H. L.; Kim, S.-H. High-Resolution Crystal Structures of Human Cyclin-Dependent Kinase 2 with and without ATP: Bound Waters and Natural Ligand as Guides for Inhibitor Design. *J. Med. Chem.* **1996,** *39* (23), 4540–4546.

[438] Hardcastle, I. R.; Golding, B. T.; Griffin, R. J. Designing Inhibitors of Cyclin-Dependent Kinases. *Annu. Rev. Pharmacol. Toxicol.* **2002,** *42* (1), 325–348.

[439] Cohen, P. Protein Kinases -- The Major Drug Targets of the Twenty-First Century? *Nat. Rev. Drug Discovery* **2002,** *1* (4), 309–315.

[440] Guha, M. Blockbuster Dreams for Pfizer's CDK Inhibitor. *Nat. Biotechnol.* **2013,** *31* (3), 187.

[441] Zuccotto, F.; Ardini, E.; Casale, E.; Angiolini, M. Through the “Gatekeeper Door”: Exploiting the Active Kinase Conformation. *J. Med. Chem.* **2009,** *53* (7), 2681–2694.

[442] Zhang, J.; Yang, P. L.; Gray, N. S. Targeting Cancer with Small Molecule Kinase Inhibitors. *Nat. Rev. Cancer* **2009,** *9* (1), 28–39.

[443] Tavares, F. X.; Boucheron, J. A.; Dickerson, S. H.; Griffin, R. J.; Preugschat, F.; Thomson, S. A.; Wang, T. Y.; Zhou, H.-Q. N-Phenyl-4-pyrazolo[1,5-b]pyridazin-3-ylpyrimidin-2-amines as Potent and Selective Inhibitors of Glycogen Synthase Kinase 3 with Good Cellular Efficacy. *J. Med. Chem.* **2004,** *47* (19), 4716–4730.

[444] Wyatt, P. G.; Woodhead, A. J.; Berdini, V.; Boulstridge, J. A.; Carr, M. G.; Cross, D. M.; Davis, D. J.; Devine, L. A.; Early, T. R.; Feltell, R. E.; Lewis, E. Jonathan; McMenamin, R. L.; Navarro, E. F.; O’Brien, M. A.; O’Reilly, M.; Reule, M.; Saxty, G.; Seavers, Lisa C. A.; Smith, D.-M.; Squires, M. S.; Trewartha, G.; Walker, M. T.; Woolford, Alison J. -A. Identification of N-(4-Piperidinyl)-4-(2,6-Dichlorobenzoylamino)-1H-Pyrazole-3-Carboxamide (AT7519), a Novel Cyclin Dependent Kinase Inhibitor Using Fragment-Based X-Ray Crystallography and Structure Based Drug Design. *J. Med. Chem.* **2008,** *51* (16), 4986–4999.

[445] Trujillo, J. I.; Kiefer, J. R.; Huang, W.; Thorarensen, A.; Xing, L.; Caspers, N. L.; Day, J. E.; Mathis, K. J.; Kretzmer, K. K.; Reitz, B. A.; Weinberg, R. A.; Stegeman, R. A.; Wrightstone, A.; Christine, L.; Compton, R.; Li, X. 2-(6-Phenyl-1H-Indazol-3-yl)-1H-Benzo[d]imidazoles: Design and Synthesis of a Potent and Isoform Selective PKC-ζ Inhibitor. *Bioorg. Med. Chem. Lett.* **2009,** *19* (3), 908–911.

[446] Aronov, A. M.; McClain, B.; Moody, C. S.; Murcko, M. A. Kinase-Likeness and Kinase-Privileged Fragments: Toward Virtual Polypharmacology. *J. Med. Chem.* **2008,** *51* (5), 1214–1222.

[447] Jones, C. D.; Andrews, D. M.; Barker, A. J.; Blades, K.; Daunt, P.; East, S.; Geh, C.; Graham, M. A.; Johnson, K. M.; Loddick, S. A. The Discovery of AZD5597, a Potent Imidazole Pyrimidine Amide CDK Inhibitor Suitable for Intravenous Dosing. *Bioorg. Med. Chem. Lett.* **2008,** *18* (24), 6369–6373.

[448] Sayle, K. L.; Bentley, J.; Boyle, F. T.; Calvert, A. H.; Cheng, Y.; Curtin, N. J.; Endicott, J. A.; Golding, B. T.; Hardcastle, I. R.; Jewsbury, P.; Mesguiche, V.; Newell, D. R.; Noble, M. E. M.; Parsons, R. J.; Pratt, D. J.; Wang, L. Z.; Griffin, R. J. Structure-Based Design of 2-Arylamino-4-Cyclohexylmethyl-5-Nitroso-6-Aminopyrimidine Inhibitors of Cyclin-Dependent Kinases 1 and 2. *Bioorg. Med. Chem. Lett.* **2003,** *13* (18), 3079–3082.

[449] Uetrecht, J. P. Reactivity and Possible Significance of Hydroxylamine and Nitroso Metabolites of Procainamide. *J. Pharmacol. Exp. Ther.* **1985,** *232* (2), 420–425.

[450] Cheng, L.; Stewart, B. J.; You, Q.; Petersen, D. R.; Ware, J. A.; Piccotti, J. R.; Kawabata, T. T.; Ju, C. Covalent Binding of the Nitroso Metabolite of Sulfamethoxazole Is Important in Induction of Drug-Specific T-Cell Responses in Vivo. *Mol. Pharmacol.* **2008,** *73* (6), 1769–1775.

[451] Simmons, D. L.; Botting, R. M.; Hla, T. Cyclooxygenase Isozymes: The Biology of Prostaglandin Synthesis and Inhibition. *Pharmacol. Rev.* **2004,** *56* (3), 387–437.

[452] Dannhardt, G.; Laufer, S. Structural Approaches to Explain the Selectivity of COX-2 Inhibitors: Is There a Common Pharmacophore? *Curr. Med. Chem.* **2000,** *7* (11), 1101–1112.

[453] Palomer, A.; Cabré, F.; Pascual, J.; Campos, J.; Trujillo, M. A.; Entrena, A.; Gallo, M. A.; García, L.; Mauleón, D.; Espinosa, A. Identification of Novel Cyclooxygenase-2 Selective Inhibitors Using Pharmacophore Models. *J. Med. Chem.* **2002,** *45* (7), 1402–1411.

[454] Michaux, C.; Leval, X. de; Julémont, F.; Dogné, J.-M.; Pirotte, B.; Durant, F. Structure-Based Pharmacophore of COX-2 Selective Inhibitors and Identification of Original Lead Compounds from 3D Database Searching Method. *Eur. J. Med. Chem.* **2006,** *41* (12), 1446–1455.

[455] Lindner, M.; Sippl, W.; Radwan, A. A. Pharmacophore Elucidation and Molecular Docking Studies on 5-Phenyl-1-(3-Pyridyl)-1H-1, 2, 4-Triazole-3-Carboxylic Acid Derivatives as COX-2 Inhibitors. *Sci. Pharm.* **2010,** *78* (2), 195–214.

[456] Topol, E. J. Failing the Public Health — Rofecoxib, Merck, and the FDA. *N. Engl. J. Med.* **2004,** *351* (17), 1707–1709.

[457] Rao, P. N. P.; Amini, M.; Li, H.; Habeeb, A. G.; Knaus, E. E. Design, Synthesis, and Biological Evaluation of 6-Substituted-3-(4-methanesulfonylphenyl)-4-phenylpyran-2-ones: A Novel Class of Diarylheterocyclic Selective Cyclooxygenase-2 Inhibitors. *J. Med. Chem.* **2003,** *46* (23), 4872–4882.

[458] Rao, P. N. P.; Amini, M.; Li, H.; Habeeb, A. G.; Knaus, E. E. 6-Alkyl, Alkoxy, or Alkylthio-Substituted 3-(4-Methanesulfonylphenyl)-4-Phenylpyran-2-ones: A Novel Class of Diarylheterocyclic Selective Cyclooxygenase-2 Inhibitors. *Bioorg. Med. Chem. Lett.* **2003,** *13* (13), 2205–2209.

[459] Costanzo, M. J.; Almond, H. R.; Hecker, L. R.; Schott, M. R.; Yabut, S. C.; Zhang, H.-C.; Andrade-Gordon, P.; Corcoran, T. W.; Giardino, E. C.; Kauffman, J. A.; Lewis, J. M.; De Garavilla, L.; Haertlein, B. J.; Maryanoff, B. E. In-Depth Study of Tripeptide-Based α-Ketoheterocycles as Inhibitors of Thrombin. Effective Utilization of the S1‘ Subsite and Its Implications to Structure-Based Drug Design1. *J. Med. Chem.* **2004,** *48* (6), 1984–2008.

[460] Tucker, T. J.; Lumma, W. C.; Lewis, S. D.; Gardell, S. J.; Lucas, B. J.; Baskin, E. P.; Woltmann, R.; Lynch, J. J.; Lyle, E. A.; Appleby, S. D.; Chen, I.-W.; Dancheck, K. B.; Vacca, J. P. Potent Noncovalent Thrombin Inhibitors That Utilize the Unique Amino Acid d-Dicyclohexylalanine in the P3 Position. Implications on Oral Bioavailability and Antithrombotic Efficacy. *J. Med. Chem.* **1997,** *40* (11), 1565–1569.

[461] Bode, W.; Turk, D.; Karshikov, A. The Refined 1.9-Å X-Ray Crystal Structure of D-Phe-Pro-Arg Chloromethylketone-Inhibited Human α-Thrombin: Structure Analysis, Overall Structure,

Electrostatic Properties, Detailed Active-Site Geometry, and Structure-Function Relationships. *Protein Sci.* **1992,** *1* (4), 426–471.

[462] Lyle, T. A.; Chen, Z.; Appleby, S. D.; Freidinger, R. M.; Gardell, S. J.; Lewis, S. D.; Li, Y.; Lyle, E. A.; Lynch, J. J.; Mulichak, A. M. Synthesis, Evaluation, and Crystallographic Analysis of L-371,912: A Potent and Selective Active-Site Thrombin Inhibitor. *Bioorg. Med. Chem. Lett.* **1997,** *7* (1), 67–72.

[463] Roth, B. L.; Sheffler, D. J.; Kroeze, W. K. Magic Shotguns Versus Magic Bullets: Selectively Non-Selective Drugs for Mood Disorders and Schizophrenia. *Nat. Rev. Drug Discovery* **2004,** *3* (4), 353–359.

[464] Peters, J.-U.; Hert, J.; Bissantz, C.; Hillebrecht, A.; Gerebtzoff, G.; Bendels, S.; Tillier, F.; Migeon, J.; Fischer, H.; Guba, W.; Kansy, M. Can We Discover Pharmacological Promiscuity Early in the Drug Discovery Process? *Drug Discovery Today* **2012,** *17* (7–8), 325–335.

[465] Leung, D.; Abbenante, G.; Fairlie, D. P. Protease Inhibitors: Current Status and Future Prospects. *J. Med. Chem.* **2000,** *43* (3), 305–341.

[466] Aronov, A. M.; Murcko, M. A. Toward a Pharmacophore for Kinase Frequent Hitters. *J. Med. Chem.* **2004,** *47* (23), 5616–5619.

[467] Chen, C.-Y.; Chang, Y.-L.; Shih, J.-Y.; Lin, J.-W.; Chen, K.-Y.; Yang, C.-H.; Yu, C.-J.; Yang, P.-C. Thymidylate Synthase and Dihydrofolate Reductase Expression in Non-Small Cell Lung Carcinoma: The Association with Treatment Efficacy of Pemetrexed. *Lung Cancer* **2011,** *74* (1), 132–138.

[468] Hopkins, A. L.; Mason, J. S.; Overington, J. P. Can We Rationally Design Promiscuous Drugs? *Curr. Opin. Struct. Biol.* **2006,** *16* (1), 127–136.

[469] Hill, E. R.; Tian, J.; Tilley, M. R.; Zhu, M. X.; Gu, H. H. Potencies of Cocaine Methiodide on Major Cocaine Targets in Mice. *PloS one* **2009,** *4* (10), e7578.

[470] Morphy, R.; Kay, C.; Rankovic, Z. From Magic Bullets to Designed Multiple Ligands. *Drug Discovery Today* **2004,** *9* (15), 641–651.

[471] Germain, P.; Staels, B.; Dacquet, C.; Spedding, M.; Laudet, V. Overview of Nomenclature of Nuclear Receptors. *Pharmacol. Rev.* **2006,** *58* (4), 685–704.

[472] Păunescu, H.; Coman, O. A.; Coman, L.; Ghiţă, I.; Georgescu, S. R.; Drăia, F.; Fulga, I. Cannabinoid System and Cyclooxygenases Inhibitors. *J. Med. Life* **2011,** *4* (1), 11–20.

[473] D'Ambra, T. E.; Estep, K. G.; Bell, M. R.; Eissenstat, M. A.; Josef, K. A.; Ward, S. J.; Haycock, D. A.; Baizman, E. R.; Casiano, F. M. Conformationally Restrained Analogs of Pravadoline: Nanomolar Potent, Enantioselective,(Aminoalkyl) Indole Agonists of the Cannabinoid Receptor. *J. Med. Chem.* **1992,** *35* (1), 124–135.

[474] Rezende, R. M.; Paiva-Lima, P.; Dos Reis, W. G.; Camêlo, V. M.; Faraco, A.; Bakhle, Y. S.; Francischi, J. N. Endogenous Opioid and Cannabinoid Mechanisms are Involved in the Analgesic Effects of Celecoxib in the Central Nervous System. *Pharmacology* **2012,** *89* (3-4), 127–136.

[475] Le Staniaszek; Norris, L. M.; Kendall, D. A.; Barrett, D. A.; Chapman, V. Effects of COX-2 Inhibition on Spinal Nociception: The Role of Endocannabinoids. *Br. J. Pharmacol.* **2010,** *160* (3), 669–676.

[476] Warden, S. J. Prophylactic Misuse and Recommended Use of Non-Steroidal Anti-Inflammatory Drugs by Athletes. *Br. J. Sports Med.* **2009,** *43* (8), 548–549.

[477] Patsos, H. A.; Hicks, D. J.; Dobson, R R H; Greenhough, A.; Woodman, N.; Lane, J. D.; Williams, A. C.; Paraskeva, C. The Endogenous Cannabinoid, Anandamide, Induces Cell

Death in Colorectal Carcinoma Cells: A Possible Role for Cyclooxygenase 2. *Gut* **2005,** *54* (12), 1741–1750.

[478] "Celecoxib (CELEBREX) und Depression", arznei-telegramm 2007, 38, 105.

[479] Lehmann, J. M.; Lenhard, J. M.; Oliver, B. B.; Ringold, G. M.; Kliewer, S. A. Peroxisome Proliferator-Activated Receptors α and γ Are Activated by Indomethacin and Other Non-Steroidal Anti-Inflammatory Drugs. *J. Biol. Chem.* **1997,** *272* (6), 3406–3410.

[480] Felts, A. S.; Siegel, B. S.; Young, S. M.; Moth, C. W.; Lybrand, T. P.; Dannenberg, A. J.; Marnett, L. J.; Subbaramaiah, K. Sulindac Derivatives that Activate the Peroxisome Proliferator-Activated Receptor γ but Lack Cyclooxygenase Inhibition. *J. Med. Chem.* **2008,** *51* (16), 4911–4919.

[481] Wick, M.; Hurteau, G.; Dessev, C.; Chan, D.; Geraci, M. W.; Winn, R. A.; Heasley, L. E.; Nemenoff, R. A. Peroxisome Proliferator-Activated receptor-γ Is a Target of Nonsteroidal Anti-Inflammatory Drugs Mediating Cyclooxygenase-Independent Inhibition of Lung Cancer Cell Growth. *Mol. Pharmacol.* **2002,** *62* (5), 1207–1214.

[482] Badawi, A. F.; Badr, M. Z. Chemoprevention of Breast Cancer by Targeting Cyclooxygenase-2 and Peroxisome Proliferator-Activated Receptor-γ (Review). *Int. J. Oncol.* **2002,** *20* (6), 1109.

[483] O'Sullivan, S. E. Cannabinoids Go Nuclear: Evidence for Activation of Peroxisome Proliferator-Activated Receptors. *Br. J. Pharmacol.* **2007,** *152* (5), 576–582.

[484] Sun, Y.; Bennett, A. Cannabinoids: A New Group of Agonists of PPARs. *PPAR Res.* **2007,** *2007*, 1–7.

[485] O'Sullivan, S. E.; Kendall, D. A. Cannabinoid Activation of Peroxisome Proliferator-Activated Receptors: Potential for Modulation of Inflammatory Disease. *Immunobiology* **2010,** *215* (8), 611–616.

[486] Scheen, A. J.; Finer, N.; Hollander, P.; Jensen, M. D.; Van Gaal, Luc F. Efficacy and Tolerability of Rimonabant in Overweight or Obese Patients with Type 2 Diabetes: A Randomised Controlled Study. *Lancet* **2006,** *368* (9548), 1660–1672.

[487] Derosa, G.; Maffioli, P. Anti-Obesity Drugs: A Review about their Effects and their Safety. *Expert Opin. Drug Saf.* **2012,** *11* (3), 459–471.

[488] Nicholson, C. D.; Jackman, S. A.; Wilke, R. The Ability of Denbufylline to Inhibit Cyclic Nucleotide Phosphodiesterase and Its Affinity for Adenosine Receptors and the Adenosine Re-Uptake Site. *Br. J. Pharmacol.* **1989,** *97* (3), 889–897.

[489] Ukena, D.; Schudt, C.; Sybrecht, G. W. Adenosine Receptor-Blocking Xanthines as Inhibitors of Phosphodiesterase Isozymes. *Biochem. Pharmacol.* **1993,** *45* (4), 847–851.

[490] Francis, S. H.; Turko, I. V.; Corbin, J. D. Cyclic Nucleotide Phosphodiesterases: Relating Structure and Function. In *Progress in Nucleic Acid Research and Molecular Biology*; Moldave, K., Ed.; Academic Press: San Diego, 2000, Bd. 65; S. 1–52.

[491] Fredholm, B. B.; Arslan, G.; Halldner, L.; Kull, B.; Schulte, G.; Wasserman, W. Structure and Function of Adenosine Receptors and Their Genes. *Naunyn-Schmiedeberg's Arch. Pharmacol.* **2000,** *362* (4-5), 364-374.

[492] Goettert, M.; Schattel, V.; Koch, P.; Merfort, I.; Laufer, S. Biological Evaluation and Structural Determinants of p38α Mitogen-Activated-Protein Kinase and c-Jun-N-Terminal Kinase 3 Inhibition by Flavonoids. *ChemBioChem* **2010,** *11* (18), 2579–2588.

[493] Pacher, P.; Nivorozhkin, A.; Szabó, C. Therapeutic Effects of Xanthine Oxidase Inhibitors: Renaissance Half a Century after the Discovery of Allopurinol. *Pharmacol. Rev.* **2006,** *58* (1), 87–114.

[494] Kou, B.; Ni, J.; Vatish, M.; Singer, Donald R. J. Xanthine Oxidase Interaction with Vascular Endothelial Growth Factor in Human Endothelial Cell Angiogenesis. *Microcirculation* **2008,** *15* (3), 251–267.

[495] Schroeter, H.; Boyd, C.; Spencer, Jeremy P. E.; Williams, R. J.; Cadenas, E.; Rice-Evans, C. MAPK Signaling in Neurodegeneration: Influences of Flavonoids and of Nitric Oxide. *Neurobiol. Aging* **2002,** *23* (5), 861–880.

[496] Simpson, E. R.; Mahendroo, M. S.; Means, G. D.; Kilgore, M. W.; Hinshelwood, M. M.; Graham-Lorence, S.; Amarneh, B.; Ito, Y.; Fisher, C. R.; Michael, M. D. Aromatase Cytochrome P450, the Enzyme Responsible for Estrogen Biosynthesis. *Endocr. Rev.* **1994,** *15* (3), 342–355.

[497] Santen, R. J.; Brodie, H.; Simpson, E. R.; Siiteri, P. K.; Brodie, A. History of Aromatase: Saga of an Important Biological Mediator and Therapeutic Target. *Endocr. Rev.* **2009,** *30* (4), 343–375.

[498] van den Beukel, I.; Dijcks, F. A.; Vanderheyden, P.; Vauquelin, G.; Oortgiesen, M. Differential Muscarinic Receptor Binding of Acetylcholinesterase Inhibitors in Rat Brain, Human Brain and Chinese Hamster Ovary Cells Expressing Human Receptors. *J. Pharmacol. Exp. Ther.* **1997,** *281* (3), 1113–1119.

[499] Benzi, G.; Moretti, A. Is There a Rationale for the Use of Acetylcholinesterase Inhibitors in the Therapy of Alzheimer's Disease? *Eur. J. Pharmacol.* **1998,** *346* (1), 1–13.

[500] Friedman, J. Cholinergic Targets for Cognitive Enhancement in Schizophrenia: Focus on Cholinesterase Inhibitors and Muscarinic Agonists. *Psychopharmacology* **2004,** *174* (1), 45-53.

[501] Petroianu, G.; Arafat, K.; Sasse, B. C.; Stark, H. Multiple Enzyme Inhibitions by Histamine H3 Receptor Antagonists as Potential Procognitive Agents. *Pharmazie in unserer Zeit* **2006,** *61* (3), 179–182.

[502] Morphy, R.; Rankovic, Z. Designed Multiple Ligands. An Emerging Drug Discovery Paradigm. *J. Med. Chem.* **2005,** *48* (21), 6523–6543.

[503] Lednicer, D. Tracing the Origins of COX-2 Inhibitors Structures. *Curr. Med. Chem.* **2002,** *9* (15), 1457–1461.

[504] Leonard, B. Sigma Receptors and Sigma Ligands: Background to a Pharmacological Enigma. *Pharmacopsychiatry* **2004,** *37* (S 3), 166–170.

[505] Glennon, R. Pharmacophore Identification for Sigma-1 (σ1) Receptor Binding: Application of the "Deconstruction - Reconstruction - Elaboration" Approach. *Mini-Rev. Med. Chem.* **2005,** *5* (10), 927–940.

[506] Maurice, T.; Su, T.-P. The Pharmacology of Sigma-1 Receptors. *Pharmacol. Ther.* **2009,** *124* (2), 195–206.

[507] Hayashi, T.; Su, T.-P. Sigma-1 Receptor Chaperones at the ER-Mitochondrion Interface Regulate Ca^{2+} Signaling and Cell Survival. *Cell* **2007,** *131* (3), 596–610.

[508] Su, T.-P.; Hayashi, T.; Maurice, T.; Buch, S.; Ruoho, A. E. The Sigma-1 Receptor Chaperone as an Inter-Organelle Signaling Modulator. *Trends Pharmacol. Sci.* **2010,** *31* (12), 557–566.

[509] Cobos, E. J.; Entrena, J. M.; Nieto, F. R.; Cendan, C. M.; Del Pozo, E. Pharmacology and Therapeutic Potential of Sigma1 Receptor Ligands. *Curr. Neuropharmacol.* **2008,** *6* (4), 344–366.

[510] Hayashi, T.; Su, T.-P. Cholesterol at the Endoplasmic Reticulum: Roles of the Sigma-1 Receptor Chaperone and Implications thereof in Human Diseases. *Subcell. Biochem.* **2010,** *51*, 381–398.

[511] Teruo, H.; Shang-Yi, T.; Tomohisa, M.; Michiko, F.; Tsung-Ping, S. Targeting Ligand-Operated Chaperone Sigma-1 Receptors in the Treatment of Neuropsychiatric Disorders. *Expert Opin. Ther. Targets* **2011,** *15* (5), 557–577.

[512] Ishikawa, M.; Hashimoto, K. The Role of Sigma-1 Receptors in the Pathophysiology of Neuropsychiatric Diseases. *J. Recept., Ligand Channel Res.* **2010,** *3*, 25–36.

[513] Ulus, I. H.; Maher, T. J.; Wurtman, R. J. Characterization of Phentermine and Related Compounds as Monoamine Oxidase (Mao) Inhibitors. *Biochem. Pharmacol.* **2000,** *59* (12), 1611–1621.

[514] Fišar, Z. Inhibition of Monoamine Oxidase Activity by Cannabinoids. *Naunyn-Schmiedeberg's Arch. Pharmacol.* **2010,** *381* (6), 563-572.

[515] Fišar, Z. Cannabinoids and Monoamine Neurotransmission with Focus on Monoamine Oxidase. *Prog. Neuro-Psychopharmacol. Biol. Psychiatry* **2012,** *38* (1), 68–77.

[516] Bailly, F.; Queffélec, C.; Mbemba, G.; Mouscadet, J.-F.; Pommery, N.; Pommery, J.; Hénichart, J.-P.; Cotelle, P. Synthesis and Biological Activities of a Series of 4,5-Diaryl-3-Hydroxy-2(5H)-Furanones. *Eur. J. Med. Chem.* **2008,** *43* (6), 1222–1229.

[517] Nakatani, K.; Nakahata, N.; Arakawa, T.; Yasuda, H.; Ohizumi, Y. Inhibition of Cyclooxygenase and Prostaglandin E_2 Synthesis by γ-Mangostin, a Xanthone Derivative in Mangosteen, in C6 Rat Glioma Cells. *Biochem. Pharmacol.* **2002,** *63* (1), 73–79.

[518] Ohishi, N.; Suzuki, T.; Ogasawara, T.; Yagi, K. Xanthone Derivatives as Inhibitors for Monoamine Oxidase. *J. Mol. Catal. B: Enzym.* **2000,** *10* (1–3), 291–294.

[519] Altiok, N.; Koyuturk, M.; Altiok, S. JNK Pathway Regulates Estradiol-Induced Apoptosis in Hormone-Dependent Human Breast Cancer Cells. *Breast Cancer Res. Treat.* **2007,** *105* (3), 247-254.

[520] Altiok, N.; Ersoz, M.; Koyuturk, M. Estradiol Induces JNK-Dependent Apoptosis in Glioblastoma Cells. *Oncol. Lett.* **2011,** *2* (6), 1281–1285.

[521] Antoon, J. W.; White, M. D.; Meacham, W. D.; Slaughter, E. M.; Muir, S. E.; Elliott, S.; Rhodes, L. V.; Ashe, H. B.; Wiese, T. E.; Smith, C. D.; Burow, M. E.; Beckman, B. S. Antiestrogenic Effects of the Novel Sphingosine Kinase-2 Inhibitor ABC294640. *Endocrinology* **2010,** *151* (11), 5124–5135.

[522] Stiles, G. L.; Caron, M. G.; Lefkowitz, R. J. Beta-Adrenergic Receptors: Biochemical Mechanisms of Physiological Regulation. *Physiol. Rev.* **1984,** *64* (2), 661–743.

[523] Williams, L. T.; Lefkowitz, R. J.; Watanabe, A. M.; Hathaway, D. R.; Besch, H. R. Thyroid Hormone Regulation of Beta-Adrenergic Receptor Number. *J. Biol. Chem.* **1977,** *252* (8), 2787–2789.

[524] Hook, Vivian Y. H.; Kindy, M.; Hook, G. Inhibitors of Cathepsin B Improve Memory and Reduce β-Amyloid in Transgenic Alzheimer Disease Mice Expressing the Wild-type, but Not the Swedish Mutant, β-Secretase Site of the Amyloid Precursor Protein. *J. Biol. Chem.* **2008,** *283* (12), 7745–7753.

[525] Sachiko Katayama; Kazutaka Shigemi; Mark A. Cline; Mitsuhiro Furuse. Clorgyline Inhibits Orexin-A-Induced Arousal in Layer-Type Chicks. *J. Vet. Med. Sci.* **2011,** *73* (4), 471–474.

[526] Small, C. J. Central Orexin A Has Site-Specific Effects on Luteinizing Hormone Release in Female Rats. *Endocrinology* **2003,** *144* (7), 3225–3236.

[527] Easton, A.; Dwyer, E.; Pfaff, D. W. Estradiol and Orexin-2 Saporin Actions on Multiple Forms of Behavioral Arousal in Female Mice. *Behav. Neurosci.* **2006,** *120* (1), 1–9.

[528] Hondo, M.; Nagai, K.; Ohno, K.; Kisanuki, Y.; Willie, J. T.; Watanabe, T.; Yanagisawa, M.; Sakurai, T. Histamine-1 Receptor Is Not Required as a Downstream Effector of Orexin-2 Receptor in Maintenance of Basal Sleep/Wake States. *Acta Physiol.* **2010,** *198* (3), 287–294.

[529] Sundvik, M.; Kudo, H.; Toivonen, P.; Rozov, S.; Chen, Y.-C.; Panula, P. The Histaminergic System Regulates Wakefulness and Orexin/Hypocretin Neuron Development via Histamine Receptor H1 in Zebrafish. *FASEB J.* **2011,** *25* (12), 4338–4347.

[530] Lin, J.-S.; Dauvilliers, Y.; Arnulf, I.; Bastuji, H.; Anaclet, C.; Parmentier, R.; Kocher, L.; Yanagisawa, M.; Lehert, P.; Ligneau, X. An Inverse Agonist of the Histamine H_3 Receptor Improves Wakefulness in Narcolepsy: Studies in Orexin$^{-/-}$ Mice and Patients. *Neurobiol. Dis.* **2008,** *30* (1), 74–83.

[531] Watson, C.; Kreuzaler, P. The Role of Cathepsins in Involution and Breast Cancer. *J. Mammary Gland. Biol. Neoplasia* **2009,** *14* (2), 171-179.

[532] Sirvent, A.; Benistant, C.; Roche, S. Cytoplasmic signalling by the c-Abl Tyrosine Kinase in Normal and Cancer Cells. *Biol. Cell* **2008,** *100* (11), 617–631.

[533] Stankovic, C. J.; Surendran, N.; Lunney, E. A.; Plummer, M. S.; Para, K. S.; Shahripour, A.; Fergus, J. H.; Marks, J. S.; Herrera, R.; Hubbell, S. E.; Humblet, C.; Saltiel, A. R.; Stewart, B. H.; Sawyer, T. K. The Role of 4-Phosphonodifluoromethyl- and 4-Phosphono-phenylalanine in the Selectivity and Cellular Uptake of SH2 Domain Ligands. *Bioorg. Med. Chem. Lett.* **1997,** *7* (14), 1909–1914.

[534] Eaton, S. R.; Cody, W. L.; Doherty, A. M.; Holland, D. R.; Panek, R. L.; Lu, G. H.; Dahring, T. K.; Rose, D. R. Design of Peptidomimetics That Inhibit the Association of Phosphatidylinositol 3-Kinase with Platelet-Derived Growth Factor-β Receptor and Possess Cellular Activity. *J. Med. Chem.* **1998,** *41* (22), 4329–4342.

[535] Al-Obeidi, F. A.; Lam, K. S. Development of Inhibitors for Protein Tyrosine Kinases. *Oncogene* **2000,** *19* (49), 5690–5701.

[536] Kraskouskaya, D.; Duodu, E.; Arpin, C. C.; Gunning, P. T. Progress Towards the Development of SH2 Domain Inhibitors. *Chem. Soc. Rev.* **2013,** *42* (8), 3337–3370.

[537] Sarma, R.; Sinha, S.; Ravikumar, M.; Kumar, M. K.; Mahmood, S. K. Pharmacophore Modeling of Diverse Classes of p38 MAP Kinase Inhibitors. *Eur. J. Med. Chem.* **2008,** *43* (12), 2870–2876.

[538] Hanson, G. J. Inhibitors of p38 Kinase. *Expert Opin. Ther. Pat.* **1997,** *7* (7), 729–733.

[539] Boehm, J. C.; Smietana, J. M.; Sorenson, M. E.; Garigipati, R. S.; Gallagher, T. F.; Sheldrake, P. L.; Bradbeer, J.; Badger, A. M.; Laydon, J. T.; Lee, J. C.; Hillegass, L. M.; Griswold, D. E.; Breton, J. J.; Chabot-Fletcher, M. C.; Adams, J. L. 1-Substituted 4-Aryl-5-pyridinylimidazoles: A New Class of Cytokine Suppressive Drugs with Low 5-Lipoxygenase and Cyclooxygenase Inhibitory Potency. *J. Med. Chem.* **1996,** *39* (20), 3929–3937.

[540] Börsch-Haubold, A. G.; Pasquet, S.; Watson, S. P. Direct Inhibition of Cyclooxygenase-1 and -2 by the Kinase Inhibitors SB 203580 and PD 98059: SB 203580 also Inhibits Thromboxane Synthase. *J. Biol. Chem.* **1998,** *273* (44), 28766–28772.

[541] Sperandio da Silva, Gilberto M.; Lima, L. M.; Fraga, Carlos A. M.; Sant'Anna, Carlos M. R.; Barreiro, E. J. The Molecular Basis for Coxib Inhibition of p38α MAP Kinase. *Bioorg. Med. Chem. Lett.* **2005,** *15* (15), 3506–3509.

[542] Badger, A. M.; Bradbeer, J. N.; Votta, B.; Lee, J. C.; Adams, J. L.; Griswold, D. E. Pharmacological Profile of SB 203580, a Selective Inhibitor of Cytokine Suppressive Binding Protein/p38 Kinase, in Animal Models of Arthritis, Bone Resorption, Endotoxin Shock and Immune Function. *J. Pharmacol. Exp. Ther.* **1996,** *279* (3), 1453–1461.

[543] Tong, L.; Pav, S.; Della White, M.; Rogers, S.; Crane, K. M.; Cywin, C. L.; Brown, M. L.; Pargellis, C. A. A Highly Specific Inhibitor of Human p38 MAP Kinase Binds in the ATP Pocket. *Nat. Struct. Biol.* **1997,** *4* (4), 311–316.

[544] Wilson, K. P.; McCaffrey, P. G.; Hsiao, K.; Pazhanisamy, S.; Galullo, V.; Bemis, G. W.; Fitzgibbon, M. J.; Caron, P. R.; Murcko, M. A.; Su, Michael S. S. The Structural Basis for the Specificity of Pyridinylimidazole Inhibitors of p38 MAP Kinase. *Chem. Biol.* **1997,** *4* (6), 423–431.

[545] UniProt Consortium. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **2007,** *35* (suppl 1), D193-D197.

[546] Foord, S. M.; Bonner, T. I.; Neubig, R. R.; Rosser, E. M.; Pin, J.-P.; Davenport, A. P.; Spedding, M.; Harmar, A. J. International Union of Pharmacology. XLVI. G Protein-Coupled Receptor List. *Pharmacol. Rev.* **2005,** *57* (2), 279–288.

[547] IUPHAR Database. *http://www.iuphar-db.org/index.jsp.*

[548] Uniprot Consortium. Protein naming guidelines. *http://www.uniprot.org/docs/nameprot.*

[549] Olson, H.; Betton, G.; Robinson, D.; Thomas, K.; Monro, A.; Kolaja, G.; Lilly, P.; Sanders, J.; Sipes, G.; Bracken, W.; Dorato, M.; van Deun, K.; Smith, P.; Berger, B.; Heller, A. Concordance of the Toxicity of Pharmaceuticals in Humans and in Animals. *Regul. Toxicol. Pharmacol.* **2000,** *32* (1), 56–67.

[550] RDKit: Open-Source Cheminformatics and Machine Learning Software. Greg Landrum. *http://www.rdkit.org/.*

Lebenslauf

Name: Elgin Sabrina Wollenhaupt

Geburtsdatum: 27.06.1986

Geburtsort: Goslar

1998 - 2005:	Abitur am Christian-von-Dohm-Gymnasium, Goslar
10/2005 – 10/2009:	Studium der Pharmazie an der Technischen Universität Braunschweig
11/2009 - 04/2010:	Erste Hälfte des Pharmaziepraktikums in der Krankenhausapotheke der Asklepios Harzkliniken, Goslar
05/2010 - 10/2010:	Zweite Hälfte des Pharmaziepraktikums in der Löwen-Apotheke-Oker, Goslar
12/2010:	Approbation als Apothekerin
08/2013:	Praktikum bei BASF SE, Ludwigshafen
01/2011 - 01/2014:	Wissenschaftliche Mitarbeiterin am Institut für Medizinische und Pharmazeutische Chemie der Technischen Universität Braunschweig; Anfertigung der vorliegenden Dissertation im Arbeitskreis von Prof. Dr. Knut Baumann; Lehre und Praktikums-Betreuung im 5. Semester (Arzneibuchanalytik)
01/2014:	Fachapothekerin für Pharmazeutische Analytik

Posterpreise:

- 8th German Conference on Chemoinformatics, Goslar, Germany (2012).
- 3rd Strasbourg Summer School on Chemoinformatics, Strasbourg, France (2012).